Tesis de Maestría

Reconocimiento Automático de Acciones Humanas a través de Interacciones entre Humanos y Objetos

Autor:

Víctor A. Escorcia Castillo

Director:

Prof. Juan Carlos Niebles Duque Ph.D.

3 de diciembre de 2013

A mi familia.
A mis amigos.
A todos los que hicieron posible este trabajo.

Índice general

Agradecimientos

" If we knew what it was we were doing, it would not be called research, would it? "

Albert Einstein.

El trabajo que se resume en esta tesis habría sido imposible sin la ayuda de mi director de tesis, Juan Carlos Niebles. Gracias por estar siempre dispuesto a resolver mis dudas y brindarme la oportunidad de conocer cómo se hace investigación. Sus cualidades como científico, su tesón por hacer las cosas de la manera correcta y su manera de explicar los temas más tenebrosos de forma simple hacen de él un gran asesor.

Un pilar tácito pero invaluable en el desarrollo de este trabajo es mi familia. Todo el reconocimiento de mi trabajo y de mi trayectoria se lo dedico a ustedes, esperando que pueda compensar mi ausencia, aún estando tan cerca. Gracias al esfuerzo de mi padres, por su apoyo incondicional. A mi madre por lidiar con mis malos hábitos y consentirme de manera especial. Gracias a todos en general, a mis hermanos, abuelos, tíos y primos por estar pendiente de mí.

Ana María ha sido la persona que más tiempo ha pasado conmigo y consecuentemente ha hecho de sumidero de muchos momentos de estrés. Gracias por ayudarme a ver la luz en la oscuridad, gracias por todo.

También quiero agradecer a todos los amigos del laboratorio del Departamento de Ingeniería Eléctrica y Electrónica, por ser esos grandes compañeros y amigos. De manera muy especial hago un tributo a mis amigos del equipo "Electrón". Espero que sigamos encontrando en el fútbol y las celebraciones una excusa para no perder contacto.

Por último, y no menos importante, quisiera mencionar a todos aquellos amigos que han estado pendiente de mi trayectoria y han mostrado un interés sincero en el desarrollo de este trabajo, Jorge L., Jorge M. y Ariel. De manera muy especial agradezco a Fabian y a Mary quienes siempre estuvieron dispuesto a darme una mano en esos días y mínutos críticos que anteceden las entregas del trabajo de investigación.

Resumen

El cuerpo de esta tesis esta enfocado en el reconocimiento de acciones humanas en videos. Para ello, se reconoce la existencia de agentes visuales que caracterizan una acción, tales como las personas y los objetos, y la dinámica temporal de los mismos a lo largo del video.

En primer lugar, se presenta una nueva metodología computacional para la representación de las interacciones dinámicas entre humanos y objetos en videos. El algoritmo propuesto captura la naturaleza dinámica de las interacciones entre humanos y objetos modelando su evolución a la largo de la duración de la acción. A partir de esta descripción de las interacciones, se demuestra como clasificar acciones muy similares basados en el contexto y la dinámica de las interacciones

Posteriormente, se analiza el desempeño de un algoritmo que capture de manera automática la dinámica de las interacciones entre humanos y objetos. Para ello, el algoritmo hace uso del contexto temporal de las actividades dentro de un esquema de aprendizaje jerárquico discriminativo. De esta forma se obtiene un conjunto compacto de sub-acciones discriminativas con un orden temporal especifico, a partir del cual se representa de manera efectiva una actividad.

Los resultados de los algoritmos y representaciones visuales de las acciones humanas propuestos se validaron de manera experimental en dos conjuntos de videos públicos. Estas metodologías presentaron un desempeño competitivo en relación con las propuestas que ostentan el estado del arte, demostrando de esta forma la ventaja de la representación dinámica de las interacciones entre humanos y objetos para la descripción de las acciones humanas.

Introducción

El campo de estudio de visión por computador tiene como objetivo el desarrollo de algoritmos y metodologías computacionales que permiten a las máquinas entender y tomar decisiones con base en la información visual, de este modo las máquinas son capaces de emular o superar las capacidades de percepción visual con las que cuenta el cerebro humano. La consecución este fin permitirá alcanzar el futuro dibujado por muchas películas de ciencia ficción en las que las personas tenemos a nuestra disposición robots asistentes que se encargan de las tareas del hogar, ayudar en la preparación de nuestra comida favorita, sugerir la forma en la que debemos vestir e incluso vestirnos, etc (Escorcia and Niebles, 2013).

Este trabajo se enfoca en una de las tareas fundamentales en el campo de visión por computador: el análisis y reconocimiento de las acciones humanas. Los algoritmos desarrollados para este propósito asocian los estímulos visuales presentes en las imágenes y videos con una categoría de actividad que se está desarrollando en la imagen, el análisis del movimiento del cuerpo humano, el análisis y resumen de una actividad muy larga *e.g.* la construcción de un puente, entre otras. Los progresos en esta área han sido notables (Aggarwal and Ryoo, 2011). Sin embargo, los algoritmos del estado del arte aún se encuentran muy lejos de emular las capacidades visuales del cerebro humano en términos del número de actividades que son reconocidas y la complejidad de los estímulos visuales, *e.g.* cambios de puntos de vista, movimiento de la cámara, oclusiones, etc.

De los múltiples acercamientos posibles para realizar la tarea de reconocimiento de actividades, el enfoque dominante en la actualidad es a través del uso de técnicas de *aprendizaje de máquina*[1] (Aggarwal and Ryoo, 2011). Estas técnicas construyen modelos que resumen la apariencia o la dinámica de las actividades humanas a partir de las características visuales de un subconjunto de imágenes. En los últimos años, la mayor parte de las representaciones de las acciones humanas se ha volcado al uso características visuales basadas en puntos de interés espacio-temporales de bajo nivel (Laptev et al., 2008),(Jiang et al., 2012) debido a su buen

[1]Traducido del término en inglés Machine Learning

desempeño para la clasificación de acciones simples, *e.g.* caminar, saludar, etc. en un amplio rango de contextos de grabación. Esta representación, acompañada del modelo de *bolsa de palabras* (BOW[2]), se ha convertido en el modelo predilecto para los recientes desarrollos en esta área de investigación. Sin embargo, la representación de las acciones a partir de esta metodología carece de una estructura o significado semántico. Con el ánimo de solventar el inherente vacío semántico en la descripción de las actividades y aprovechando los avances alcanzados en el área de reconocimiento y seguimiento de objetos (Everingham et al., 2010),(Yang et al., 2011), este trabajo presenta una descripción de las acciones humanas en términos de las interacciones entre humanos y objetos. Es decir, las características visuales empleadas en este trabajo hacen referencia a la descripción explícita de las interacciones entre los agentes de la actividad, *i.e.* una persona y un objeto. El uso del contexto mutuo entre humanos y objetos permite distinguir actividades y acciones con patrones de movimiento similares como contestar el teléfono y tomar café. Al mismo tiempo que, una adecuada descripción de las interacciones permite distinguir acciones con patrones de movimiento similar que involucran el mismo objeto *e.g.* llamar por teléfono y contestar el teléfono.

El enfoque metodológico de este trabajo se concentra en el reconocimiento de actividades humanas a partir de la descripción espacio-temporal de las interacciones entre humanos y objetos (H&O). Sin embargo, es importante destacar que el modelado de las interacciones H&O no es específico al reconocimiento de actividades. Una de las áreas del conocimiento que se beneficia de las observaciones realizadas en este trabajo es la róbotica, donde se espera que la descripción de las interacciones H&O permita impulsar el desarrollo de robots asistenciales (Koppula et al., 2013). Estos agentes robóticos ofrecerán llenar su taza de café cuando detecten que usted halla acabado, con base en la información visual asociada con los movimientos particulares de la acción de beber y sus gestos.

Las metas alcanzadas con este trabajo son:

- Diseño e implementación de un descriptor visual para describir las interacciones entre humanos y objetos.
- Construcción de una representación de las acciones humanas basada en las interacciones entre humanos y objetos.
- Implementación y Evaluación de un algoritmo de aprendizaje de máquina para el reconocimiento de acciones humanas a partir de las interacciones entre humanos y objetos

Las contribuciones de este trabajo en el área de reconocimiento de acciones

[2]Acrónimo tomado del termino en inglés, Bag of Words

humanas son:

- Describir las acciones humanas en términos de la dinámica de las interacciones entre humanos y objetos.
- Mostrar que el modelado dinámico de las interacciones entre humanos y objetos puede ser utilizado para mejorar el reconocimiento de acciones humanas con un contexto humano-objeto idéntico.
- Analizar el desempeño de un algoritmo que define de manera automática la estructura temporal de las actividades en el contexto de interacciones entre humanos y objetos.

Este trabajo está organizado de la siguiente forma, en el capítulo 1 se analiza una representación de las acciones humanas en términos de las interacciones entre humanos y objetos. Posteriormente, en el Capítulo 2 se analiza un algoritmo de descomposición de las actividades en segmentos temporales discriminativos que capturan la dinámica y estructura de la actividad en el contexto de interacciones entre humanos y objetos. Finalmente, se presentan las conclusiones alcanzadas a través de este trabajo y los trabajos futuros que se desprenden del mismo.

Capítulo 1

Clasificación de Acciones Humanas mediante Interacciones Espacio-Temporales entre Humanos y Objetos

Los algoritmos actuales de reconocimiento de acciones humanas en videos pueden lograr resultados prometedores de clasificación en conjuntos de videos públicos provenientes de diferentes contextos, *e.g.* películas, videos de internet, etc. (Laptev et al., 2008),(Shah, 2009),(Rodriguez et al., 2008). Sin embargo, todavía es un desafío para la mayoría de los algoritmos generar descripciones semánticas o lograr una comprehensión detallada de la información visual, debido a la amplia brecha entre las representaciones visuales y la información semántica de alto nivel relacionada con los objetos, sus partes y propiedades. El algoritmo estándar de facto empleado para el reconocimiento de acciones usa una representación de *bolsa de palabras* de puntos de interés espacio-temporales (Dollar et al., 2005), (Laptev et al., 2008). Esta metodología computacional es capaz de categorizar acciones simples, *e.g.* ponerse de pie o dar la mano, con un éxito moderado, pero tienen problemas capturando características semánticas tales como objetos involucrados en una acción o las relaciones entre objetos y actores. Este capítulo se concentra en representaciones de acciones humanas que capturen características semánticas, con el fin de reducir la escasez de descripciones de este tipo inherentes en muchos enfoques tradicionales.

Figura 1.1: Los objetos juegan un rol importante para describir y categorizar acciones humanas, puesto que proporcionan información relevante acerca de las acciones que se están desarrollando. Sin embargo, reconocer los objetos involucrados en un evento puede ser insuficiente para distinguir las acciones humanas. En el ejemplo visualizado arriba, no es suficiente reconocer un teléfono para diferenciar entre contestar el teléfono (fila de arriba) y marcar el teléfono (fila de abajo). Además, modelar su configuración espacial relativa general no proporciona poder de discriminación, en tanto que el objeto esta siempre ubicado en configuraciones similares con respecto al actor. Con el fin de discriminar estas acciones, es crucial modelar cómo las relaciones humano-objeto cambian a través del tiempo.

Una dirección prometedora para aumentar el nivel de descripción semántica y que puede ser usada para la comprehensión de las acciones es el modelado de las interacciones entre humanos y objetos. Las interacciones humano-objeto son una característica poderosa que proporciona información contextual acerca de actores y objetos. Además, ésta ha mostrado ser crítica para el reconocimiento exitoso de acciones, humanos y objetos en imágenes estáticas (Yao and Fei-Fei, 2010b). Los métodos actuales para modelar interacciones en video se enfocan en capturar el contexto de ocurrencia entre los objetos y las acciones, así como las ubicaciones espaciales relativas entre objetos y actores (Gupta and Davis, 2007), (Prest et al., 2012). Estas características pueden ser suficientes cuando cada acción de interés implica un objeto diferente, *e.g.* contestar el teléfono o beber, o cuando las acciones que involucran un mismo objeto adoptan posiciones relativas particulares entre el objeto y la persona que lo manipula, *e.g.* ejecutar el servicio o realizar un golpe directo en el tenis. Sin embargo, usualmente el interés se concentra en acciones que involucran el mismo objeto como las que se aprecian en la figura 1.1, las cuales

solo pueden ser diferenciados modelando cómo el objeto se relaciona con el actor a través del tiempo. Infortunadamente, los algoritmos actuales son incapaces de codificar esta información. En este capítulo se aborda este tema introduciendo una representación que codifica las características relacionadas con las interacciones espacio-temporales entre humanos y objetos en video. El algoritmo presentado combina información acerca del objeto y actor, su posición relativa y la evolución de la interacción a través del tiempo. Integrando estas características, el algoritmo es capaz de capturar diferencias sutiles en acciones que solo difieren en la evolución temporal de la interacción humano-objeto, tales como llamar y contestar el teléfono. Adicionalmente, la metodología computacional propuesta produce una representación semántica descriptiva, de la cual se pueden aprender características significativas que mejoran el rendimiento del reconocimiento de acciones.

Este capítulo esta organizado de la siguiente forma, en la Sección 1.1 se resume el trabajo previo relacionado con el reconocimiento de acciones humanas involucrando interacciones entre humanos y objetos. Luego, la Sección 1.2 describe la representación propuesta de las interacciones espacio-temporales de humanos y objetos. Posteriormente, la Sección 1.3 presenta la metodología computacional propuesta para el reconocimiento de acciones humanas con base en la representación propuesta. Finalmente, en la Sección 1.4 se presenta la validación experimental de la representación propuesta.

Una versión de este trabajo fue presentado en "IEEE International Conference on Computer Vision"(Escorcia and Niebles, 2013).

1.1. Trabajo Previo

En esta sección, se presentan los trabajos más relevantes en reconocimiento de acciones humanas en imágenes estáticas y videos. Para una mayor cobertura de los temas de interés se recomiendan los recientes estudios acerca del análisis de actividades (Aggarwal and Ryoo, 2011) y clasificación de objetos (Everingham et al., 2010).

1.1.1. Reconocimiento de acciones humanas en imágenes estáticas empleando descripciones densas

El esquema de clasificación de imágenes estándar basado en diccionarios de parches locales (Everingham et al., 2010) acompañados con un esquema de agrupamiento espacial (Lazebnik et al., 2006) puede proporcionar estimaciones acep-

tables de las categorías de acciones en ciertos contextos (Delaitre et al., 2010). Sin embargo, esta representación estadística es menos adecuada en escenarios de categorización de acciones muy similares (Yao and Fei-Fei, 2010a). En esta situación, Yao y Fei-Fei demostraron que las estructuras de extracción discriminativas, denominadas *grouplets*, son más convenientes que el agrupamiento de parches locales (Yao and Fei-Fei, 2010a). Otra representación densa basada en patrones aleatorios discriminativos con propiedades similares ostenta el estado del arte actual en reconocimiento de acciones en imágenes estáticas (Yao et al., 2011b).

Todos los enfoques anteriores carecen de una descripción semántica de la actividad humana debido al uso de parches locales como símbolos atómicos para la descripción de las acciones.

1.1.2. Reconocimiento de acciones humanas en imágenes estáticas con base en descripciones semánticas

Recientemente algunos trabajos han considerado el uso de características semánticas para la categorización de las acciones humanas en imágenes estáticas. Yao *et al.* usan un espacio semántico de atributos de acciones, objetos y poses humanas; las imágenes son proyectadas en el espacio para obtener una categorización y representación significativa (Yao et al., 2011a). De forma similar, Sadeghi y Farhadi (Sadeghi and Farhadi, 2011) combinan frases visuales y modelos de objetos únicos para mejorar la detección de ambas entidades. Una frase visual es un objeto o un grupo de objetos asociados con un concepto semántico, por ejemplo una persona con una botella. Estas frases modelan de manera directa la descripción visual de las interacciones entre objetos. Desafortunadamente, usualmente se requieren datos de entrenamiento altamente supervisados con el fin de construir representaciones significativamente semánticas.

Otros métodos se apoyan en el uso de poses discriminativas del cuerpo humano para categorizar las acciones humanas. Yao y Fei-Fei representan la pose humana a través de esqueletos, los cuales combinados con grafos 2.5D permiten ser flexibles ante rotaciones fuera de plano (Yao and Fei-Fei, 2012). Asimismo, algunos trabajos combinan la información de poses del cuerpo humano con el contexto de la escena. Gupta *et. al* categorizan un pequeño conjunto de acciones deportivas usando una estructura bayesiana que toma en cuenta las relaciones entre las poses humanas, los objetos y el contexto de la escena (Gupta et al., 2009). Yao y Fei-Fei (Yao and Fei-Fei, 2010b) propusieron un modelo de *campos aleatorios condicionales*, CRF[1], estándar para inferir posiciones de partes del cuerpo y de los objetos, apro-

[1]Acrónimo tomado del término en inglés, "Conditional Random Field"

vechando el contexto mutuo entre ellos. Recientemente, Desai y Ramanan (Desai and Ramanan, 2012) combinan el concepto de frases visuales discriminativas para categorizar acciones humanas e inferir poses del cuerpo humano. La principal desventaja de todas estas metodologías es que la estimación de la pose humana es todavía una tarea muy desafiante para los algoritmos actuales, por lo tanto estos trabajos requieren una gran cantidad de supervisión para reconocer las acciones humanas.

En esta dirección y aún más relacionado con el objetivo de este capítulo, algunos investigadores han propuesto la introducción de características de interacción humano-objeto sin modelar la pose del cuerpo humano directamente (Desai et al., 2010). Ésta es una metodología intermedia que aprovecha las relaciones entre actores y objetos involucradas en una acción sin requerir la exigente supervisión de la la localización de las partes del cuerpo humano. A diferencia del trabajo propuesto en este capítulo, esta metodología se desarrolló para la interpretación de imágenes estáticas y su extensión al dominio de videos no es directa.

1.1.3. Reconocimiento de acciones humanas en videos con base en descripciones semánticas

Recientemente un conjunto de trabajos han comenzado a incorporar características semánticas distintivas para el reconocimiento de acciones en videos.

Algunos autores inspirados por los trabajos realizados en el reconocimiento y categorización de objetos y escenas proponen el uso de representaciones intermedias distintivas. Éstas pueden ser vistas como espacios semánticos en los cuales se pueden proyectar videos y obtener descripciones significativas. Las opciones de espacios con rendimientos prometedores son los atributos (Liu et al., 2011) o detectores de acciones simples agrupados en un banco de acciones (Sadanand and Corso, 2012). Desafortunadamente, de manera similar a los acercamientos basados en imágenes estáticas, se requieren un conjunto de datos de entrenamiento altamente supervisados con el fin de construir estas representaciones significativamente semánticas.

Otros métodos adoptan representaciones enfocadas en la aparición de las personas en los videos para codificar características más relevantes a las actividades de interés. Usualmente, estos algoritmos emplean esquemas de seguimiento como una etapa de pre-procesamiento con el fin de localizar el actor en la escena. Posteriormente, se representan los patrones visuales y de movimiento usando características locales de video que permiten categorizar las acciones (Patron-Perez et al., 2012). Alternativamente, Lan *et. al* plantearon una representación en la cual la locali-

zación del actor es una variable latente inferida a través de un grafo de manera discriminativa (Lan et al., 2011). Esta representación ofrece una descripción más detallada de los componentes de la acción con un considerable aumento en la complejidad del modelo. Inspirados en estos trabajos, la metodología computacional propuesta en este capítulo hace uso de esquemas de detección y seguimiento visual de personas y objetos. Sin embargo, nos ocupamos de la descripción de las acciones de una manera diferente a la presentada por estos trabajos.

En esta dirección y más cercanamente relacionado con el esquema propuesto en este capítulo, algunos investigadores han propuesto la introducción de características de interacción humano-objeto, es decir a las relaciones entre actores y objetos involucrados en una acción. Algunos autores (Gupta and Davis, 2007),(Prest et al., 2013) han hecho uso de las interacciones entre humanos y objetos para el reconocimiento de acciones. Sin embargo, estas representaciones, inspiradas en los trabajos realizados en imágenes estáticas, se enfocan en las relaciones espaciales entre el humano y un objeto, por lo cual carecen de la habilidad para capturar la evolución temporal de la interacción humano-objeto a través del tiempo.

La representación presentada en este capítulo se concentra en incorporar características relacionadas a la dinámica espacio-temporal de interacciones humano-objeto en video. Por lo tanto, este trabajo también está relacionado con los métodos que codifican estructuras temporales o evolución temporal de características para reconocimiento de actividades. Esto incluye métodos que modelan relaciones espacio-temporal de características visuales, tales como segmentaciones jerárquicas espacio-temporales codificadas en estructuras gráficas (Brendel and Todorovic, 2011), características de bajo nivel agrupadas en grillas temporales (Niebles et al., 2010),(Gaidon et al., 2011), o espacio-temporales (Laptev et al., 2008), así como modelos secuenciales tradicionales tales como HMM (Ikizler and Forsyth, 2008), HCRF (Wang et al., 2006), (Tang et al., 2012a) o DBN (Laxton et al., 2007). Todos estos son revisados con más detenimiento en la sección 2.1.

1.2. Representación espacio-temporal de las interacciones entre humanos y objetos

En esta sección se expone la representación para interacciones humano-objeto en videos.

La representación propuesta captura características relativas a relaciones entre humanos y objetos, la cual se ilustra en la figura 1.2. Dada la posición de un humano $H_i^{1:T}$ y de un objeto $O_j^{1:T}$ en una secuencia de video de longitud T, el

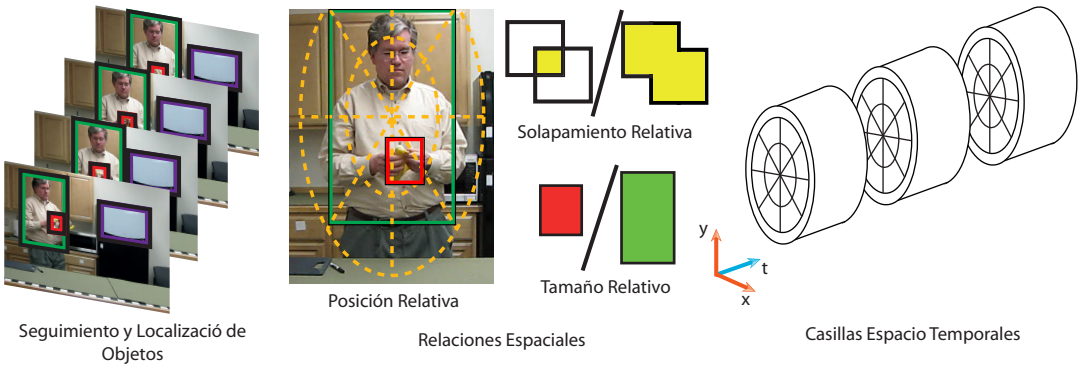


Figura 1.2: Este descriptor codifica la dinámica de interacciones espacio-temporales humanas en videos. El descriptor trabaja en pares de trayectorias (primera columna) y computa características (segunda y tercera columna) relacionadas con : la posición relativa del objeto con respecto al humano, solapamiento entre humano y objeto, y sus tamaños relativos. Con el fin de capturar cómo la interacción entre humano y objeto evoluciona a través del tiempo, este método integra estas características en intervalos de tiempo. Como ejemplo, se ilustra el caso de 3 intervalos sin solapamiento (cuarta columna), la cual define una segmentación espacio-temporal donde se integra la información de posición relativa.

objetivo de la representación propuesta es codificar la información acerca de cómo evoluciona la interacción entre el humano y el objeto a través del tiempo. Como lo muestran los experimentos, describir la evolución de las interacciones ayuda a discriminar acciones que: (1) involucran objetos similares, donde la ocurrencia de los objetos no es una pista suficiente para discriminación; (2) involucran objetos y humanos que mantienen relaciones espaciales similares, donde la medida global de la posición y tamaños relativos no es una característica distintiva; (3) sólo pueden ser discriminadas por analizar el aspecto temporal de las relaciones entre objeto y humano.

Para codificar la evolución temporal de la interacción, se integra información acerca de la posición relativa, tamaños de los objetos y humanos en casillas espacio-temporales. Intuitivamente, la cuantización temporal proporciona robustez ante el ruido y varianza al interior de los ejemplos asociados con la misma acción, con un modelo dinámico continuo, en lugar de modelar de forma cinemática la interacción.

En la práctica, esta representación adiciona las siguientes características de cada frame t en la secuencia.

Posición Relativa ϕ_l^t

Codifica la posición espacial relativa del objeto O_j^t con respecto al humano H_i^t en el frame t. Para ello se codifica esta relación usando casillas radiales dentro de una elipse en cada frame del video tal y como se ilustra en la figura 1.2. Vale la pena mencionar que se adopta una estrategia de *votación suave*, donde los objetos emiten votos con pesos que son inversamente proporcionales a la distancia de su centroide hasta el centro de la casilla. La estrategia de *votación suave* permite lidiar con la incertidumbre en la posición y forma del objeto. El uso de una casilla en forma elíptica permite tener en cuenta la relación de aspecto del cuerpo humano y suaviza la ubicación por medio de recuadros de la persona. Como se ilustra en la figura 1.2, en la práctica se usan 8 divisiones angulares y 2 divisiones radiales para caracterizar interacciones finas. Asimismo, se emplea 1 casilla, representada con el radio exterior más grande, para establecer la posición del objeto cuando no esta en contacto con la persona. Esto produce un descriptor de 17 dimensiones.

Tamaño relativo ϕ_r^t:

Codifica la relación del área en píxeles entre las ventanas del humano y del objeto, es decir $\phi_r^t = |O_j^t|/|H_i^t|$. Esta característica es útil para definir implícitamente restricciones acerca del tamaño del objeto, por ejemplo la copa es más pequeña que el humano en acciones de beber.

Solapamiento Relativo ϕ_o^t

Esta característica es computada como el área de intersección entre las ventanas del humano y del objeto. De esta forma, se combina información acerca del tamaño y la distancia del objeto con respecto al humano, en un único escalar.

Las características anteriores ϕ_l^t, ϕ_r^t y ϕ_o^t en cada frame t en el intervalo de tiempo $t = [1, T]$. Con el fin de codificar cómo estas características evolucionan a través del tiempo, éstas son integradas en muchos intervalos de tiempo. En general, se define un conjunto $\mathcal{V}$ de V intervalos de tiempo $I_v = [t_v^{start}, t_v^{end}]$. Cada intervalo v está asociado con un vector de características $\Phi^v = [\Phi_l^v, \Phi_o^v, \Phi_r^v]$, en el cual se integran las características como se muestra a continuación:

$$\Phi_l^v = \frac{1}{t_v^{end} - t_v^{start} + 1} \sum_{t \in I_v} \phi_l^t \tag{1.1}$$

$$\Phi_o^v = \left[\begin{array}{ccc} \text{máx}_{t \in I_v}\, \phi_o^t & \text{mín}_{t \in I_v}\, \phi_o^t & \overline{\phi_o^t} \end{array} \right] \tag{1.2}$$

$$\Phi_r^v = \left[\begin{array}{ccc} \text{máx}_{t \in I_v}\, \phi_r^t & \text{mín}_{t \in I_v}\, \phi_r^t & \overline{\phi_r^t} \end{array} \right] \tag{1.3}$$

Finalmente, los descriptores extraídos se concatenan en cada intervalo conformando un único vector que describe las interacciones de toda la secuencia de video.

$$\Phi(H_i, O_j) = \left[\Phi^1, \Phi^2, \ldots, \Phi^V \right] \tag{1.4}$$

La representación de las interacciones propuesta puede ser usada con muchas opciones del conjunto $\mathcal{V}$. Algunas opciones naturales son: (a) dividir un video en intervalos temporales de igual longitud sin solapamiento (figura 1.2), (b) pirámide temporal con segmentos temporales en múltiples escalas de tiempo. En esta investigación se estudian ambas opciones experimentalmente, en la evaluación empírica de la sección 1.4.

La figura 1.3 visualiza las características computadas para dos videos de ejemplos de *hacer una llamada telefónica* y *contestar una llamada telefónica*. Se nota que el uso de múltiples intervalos de tiempo (columna derecha) produce descriptores con más alto poder de discriminación en comparación a la integración global (columna central).

1.3. Clasificación de acciones

En esta sección, se describe cómo puede ser empleada la descripción de las interacciones humano-objeto propuesta para la representación de las acciones humanas en videos. A continuación, se presenta la metodología empleada para el entrenamiento de los clasificadores para cada acción de interés a partir de un conjunto de videos. Asimismo, se describe cómo realizar el reconocimiento de acciones en nuevas secuencias de video. La figura 1.4 resume el esquema de reconocimiento de acciones propuesto y la metodología algorítmica necesaria para su desarrollo.

1.3.1. Entrenamiento

El objetivo del entrenamiento es aprender a clasificar cada acción de interés. Para ello, se requiere de un conjunto de videos que contengan ejemplos de la acción que se desea reconocer. Estos videos son anotados con: la secuencia temporal donde

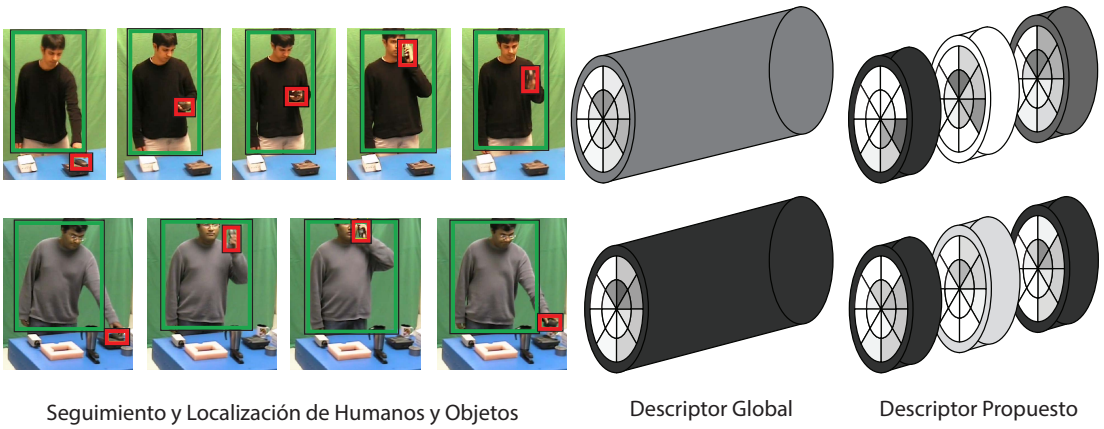


Figura 1.3: Interacciones espacio temporales entre humanos y objetos. Dado una trayectoria de objetos y humanos en un video de entrada (primera columna), el descriptor computa características relativas a las relaciones entre objeto y actor. Los métodos en comparación, integran estas características en un único segmento espacio temporal (segunda columna), ignorando toda la información temporal relativa a la interacción. El descriptor aquí descrito, integra esta información en múltiples intervalos de tiempo separadamente. En este ejemplo, el descriptor propio (tercera columna) integra información en 3 intervalos de igual longitud sin solapamiento, que cubren la secuencia completa. De esta forma, el algoritmo toma ventaja de la evolución temporal de la interacción humano-objeto, proporcionando una mejora de rendimiento en la tarea de reconocimiento de acciones humanas.

ocurre la acción, y recuadros que determinan la ubicación del humano y el objeto en al menos un frame.

Como primera etapa de pre-procesamiento, el algoritmo propuesto requiere las ubicaciones espacio temporales de la persona y el objeto a lo largo de la subsecuencia temporal que contiene la acción. Estas ubicaciones pueden ser obtenidas mediante una combinación de algoritmos de seguimiento con base en características de bajo nivel y detectores de objetos como en (Prest et al., 2013). El objetivo de esta etapa es proveer los recuadros del objeto y la persona en cada frame necesarios para capturar la dinámica de la interacción entre humano y objeto. En la práctica, se utilizó un algoritmo simple de seguimiento de bajo nivel basado en plantillas de correlación a partir de los recuadros suministrados. Este esquema de seguimiento simple puede fallar en algunos videos, por lo que manualmente se añaden algunas anotaciones hasta que el resultado del algoritmo de seguimiento sea aceptable. En otras palabras, se decidió evaluar a la representación de las acciones humanas con recuadros de gran calidad con el fin de aislar los efectos de las fallas producidos por un inadecuado seguimiento. De esta forma, la evaluación se enfoca en el poder discriminativo del descriptor propuesto.

A partir del conjunto de ubicaciones espacio temporales del humano y el objeto encontrados por el esquema de seguimiento, se computan los descriptores para todas las parejas de humano-objeto en el conjunto de entrenamiento. Para llevar a cabo esta tarea, se emplean las características introducidas en la sección 1.2, las cuales capturan la información sobre la localización del humano-objeto, sus tamaños y la evolución temporal de estas relaciones.

Finalmente, se entrena un clasificador basado en algoritmos de aprendizaje de máquina y un conjunto de videos de entrenamiento etiquetados al nivel de la categoría de acción que se desarrolla a lo largo del mismo. En la práctica, los ejemplos positivos empleados corresponden con los descriptores computados sobre las parejas humano-objeto asociadas con la acción de interés, mientras que los descriptores de cualquier pareja humano-objeto con una etiqueta diferente a la de la acción de interés son considerados como negativos. Para la clasificación de las acciones humanas se entrenó un clasificador discriminativo que determina la frontera de decisión asociada con la categoría de interés en el espacio de características definido por las interacciones humano-objeto. En la práctica, se empleó un clasificador lineal basado en *máquinas de soporte vectorial*, SVM[2].

1.3.2. Reconocimiento

Durante la etapa de reconocimiento, el algoritmo localiza y reconoce las acciones humanas en una nueva secuencia. De manera similar al entrenamiento, se utiliza una etapa de pre-procesamiento para conseguir el conjunto de localizaciones espacio-temporales de la persona y el objeto. Este proceso tiende a producir un conjunto numeroso de localizaciones, por lo que la tarea de nuestro clasificador consiste en discriminar cual pareja humano-objeto verdaderamente corresponde a cada acción de interés. Para ello, se conforman parejas de candidatos humano-objeto agrupando los conjuntos de ubicaciones que son cercanos en espacio y tiempo. Para cada pareja candidata, se computa el descriptor correspondiente tal y como se mostró en la sección 1.2. Finalmente, a partir del descriptor asociado con cada pareja se calcula su nivel de certeza con base en el modelo discriminativo de clasificación. Una pareja se declara como positiva, si su nivel certeza es mayor que el umbral establecido para la acción. En un escenario de múltiples clases, se prueba cada pareja humano-objeto contra todos los clasificadores de acciones humanas y la decisión de clasificación corresponde con la de la acción con el nivel de certeza más alto.

[2]Acrónimo tomado del término en inglés “Support Vector Machine”

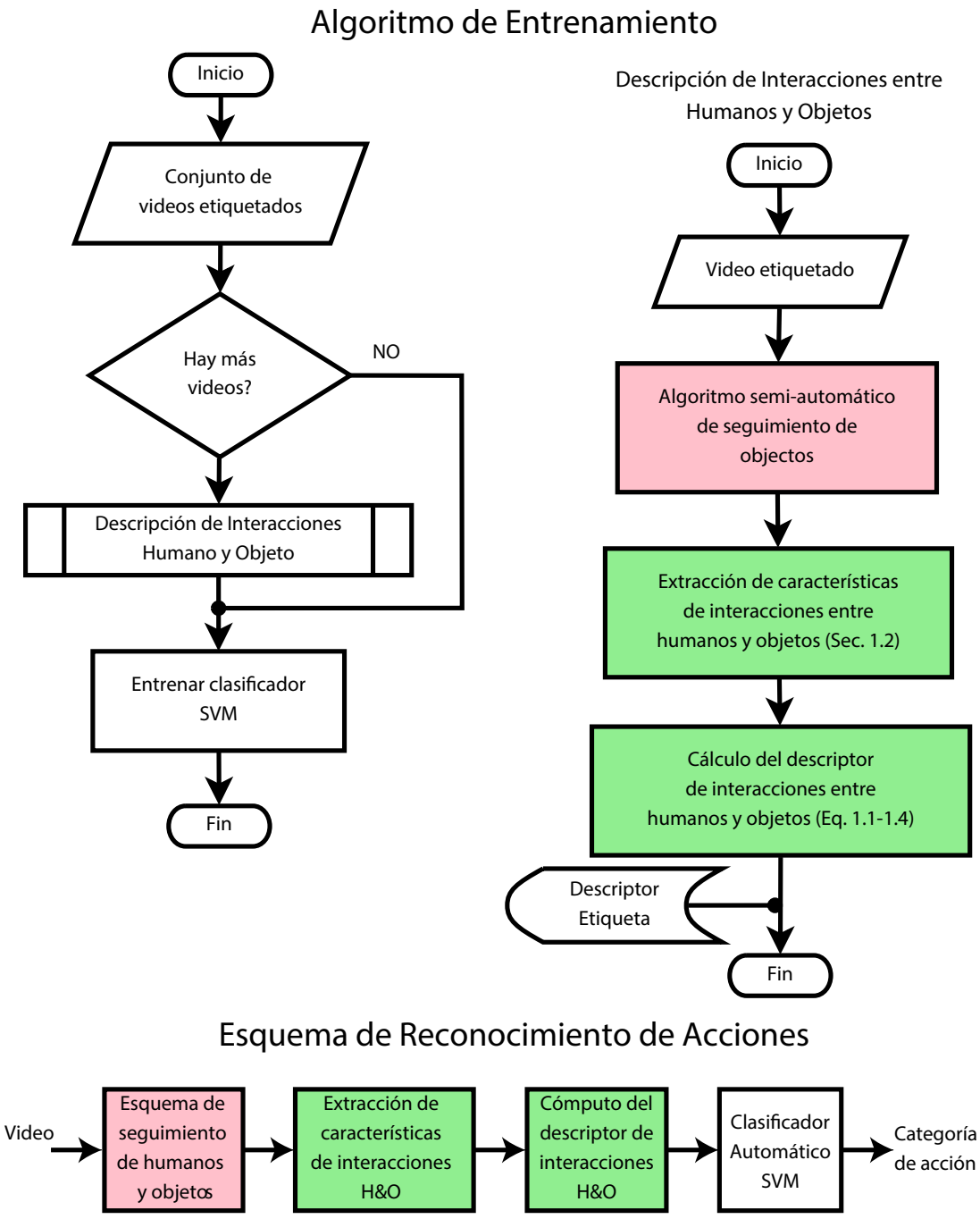


Figura 1.4: Diagrama de bloques del esquema de reconocimiento de acciones propuesto junto con el diagrama de flujo del algoritmo de la etapa de entrenamiento. Los detalles específicos de cada bloque se aprecian en la sección 1.2 y 1.3

1.4. Resultados Experimentales

Para validar las características y la representación propuesta para la descripción de las interacciones, se evaluó el algoritmo de reconocimiento de acciones humanas para la tarea de clasificación usando dos conjuntos de videos públicos (Gupta and Davis, 2007),(Messing et al., 2009).

1.4.1. Conjunto de videos de Gupta

En primer lugar, se evaluó el esquema propuesto para el reconocimiento de acciones humanas usando el conjunto de videos de acciones humanas con múltiples clases propuesto en (Gupta and Davis, 2007). Este conjunto de videos contiene unos 10 actores desarrollando 6 acciones que involucran 4 clases de objetos para un total de 54 videos. Las acciones en este conjunto de videos son: beber de una taza, rociar de una botella, contestar una llamada telefónica, hacer una llamada telefónica, servir de una taza y encender una linterna. Los videos fueron tomados en un laboratorio, usando una cámara estática en una escena con un fondo verde y objetos blancos.

Para evaluar el poder discriminativo de la representación propuesta, se utilizó un esquema de validación cruzada con cuatro conjuntos. Los clasificadores de las acciones humanas se entrenaron empleando las características de interacciones espacio temporales descritas anteriormente junto con un vector de activación booleana que captura la información contextual entre las acciones y los objetos de forma similar a como se propuso en (Prest et al., 2013). En todos los casos, se entrenaron clasificadores binarios usando los ejemplos de otras clases como negativos. Durante la etapa de reconocimiento, todas las parejas humanos-objetos encontradas en un video son examinadas por cada uno de los clasificadores binarios y la predicción de la acción asociada con el video corresponde con la clasificador con el mayor nivel de certeza sobre todas las parejas.

Con el ánimo de enfocar la evaluación en el poder discriminativo de de las características y representaciones propuestas, se usó el mismo conjunto de recuadros espacio temporales de humanos y objetos cuando se compararon múltiples métodos. La figura 1.5(a) resume los resultados de la evaluación cuantitativa del descriptor propuesto en término de una matriz de confusión, cuyo promedio en la diagonal es de 96.3 %. En esta figura también se compara el desempeño del descriptor propuesto contra una implementación propia del algoritmo que ostenta el estado del arte en la descripción de las interacciones entre humanos y objetos de (Prest et al., 2013)). Este algoritmo calcula las interacciones y relaciones en un

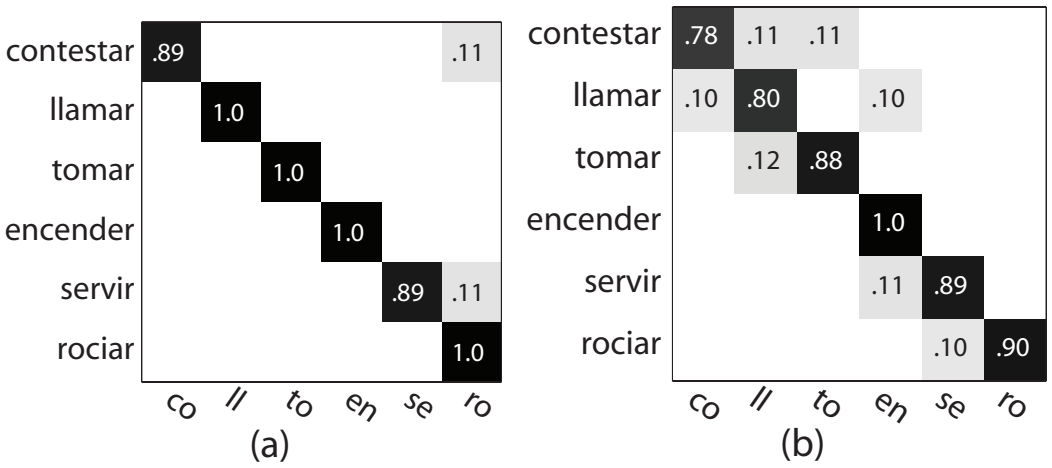


Figura 1.5: Matrices de confusión en el conjunto de videos de Gupta. (a) El descriptor propuesto que captura las interacciones espacio temporales entre humano y objeto, (b) interacción humano-objeto descrito en (Prest et al., 2013).

único intervalo temporal y no captura la evolución temporal de la interacción. La matriz de confusión asociada con este algoritmo se presenta en 1.5(b), con un promedio en la diagonal de 87.5 %. Vale la pena destacar que el descriptor propuesto provee una mayor discriminación, especialmente cuando las acciones de interés involucran objetos similares y relaciones espaciales que solo pueden ser discriminadas usando el aspecto temporal de la interacción.

Adicionalmente, se evaluó la contribución de cada componente del descriptor propuesto hasta la presentación de reconocimiento final. La tabla 1.1 resume las comparaciones cuantitativas para cada componente empleando dos métricas de

Tabla 1.1: Evaluación cuantitativa de los diferentes componentes del descriptor propuesto en el conjunto de videos de Gupta (a) Acumula Φ_l sobre un único intervalo de tiempo que cubre toda la secuencia. (b) Combinación de (Φ_l, Φ_s, Φ_o) sobre un único intervalo de tiempo que cubre toda la secuencia. (c) y (d) agregan las características sobre 3 intervalos temporales no solapados de longitud uniforme que cubren toda la duración de la secuencia.

Método	Precisión	AP promedio
Descriptor HOI Global (Prest et al., 2013)	87.5 %	91.9 %
(a) Posición Relativa Global (Φ_l)	82.5 %	80.2 %
(b) Interacciones Globales (Φ_l, Φ_s, Φ_o)	90.1 %	83.5 %
(c) Posición Relativa Espacio-Temporal	88.5 %	87.3 %
(d) Interacciones Espacio-Temporales	**96.3 %**	**93.2 %**

Tabla 1.2: Comparación experimental de diferentes alternativas estructurales para el conjunto $\mathcal{V}$.

Alternativas de estructuras temporales $\mathcal{V}$	Precisión	AP promedio
Pirámide Espacio-Temporal, 3-niveles	85.2 %	**95.5** %
Segmentos planos, 1 intervalo global	90.1 %	83.5 %
Segmentos planos, 3 intervalos no-solapados	**96.3 %**	93.2 %

desempeño: precisión, la cual corresponde al promedio de la diagonal de la matriz de confusión en el escenario de evaluación de múltiples clases; y el *AP promedio*, el cual corresponde al promedio de los valores AP[3] de cada clasificación de binario. Nótese que cada característica es complementaria y que la representación propuesta alcance el mejor desempeño combinando la posición relativa, el tamaño relativo, el solapamiento enmarcados dentro de una descripción espacio temporal. La tabla 1.1 también compara el descriptor propuesto contra una versión de referencia que acumula las características en único intervalo temporal, es decir, de manera similar a la representación propuesta por (Prest et al., 2013). Vale la pena aclarar, que los resultados obtenidos con la implementación propia del algoritmo (Prest et al., 2013) usaron el mismo conjunto de localizaciones de humanos y objetos que el descriptor propuesto. De esta forma, se garantiza una comparación justa y válida de las bondades de ambas descripciones.

Complementando el análisis de la representación de las interacciones entre humanos y objetos a través de las características propuestas, se estudió el efecto de diferentes alternativas para el conjunto $\mathcal{V}$. La tabla 1.2 resume los resultados encontrados para tipos de arreglos temporales. En ésta se aprecia que el uso de una estructura espacio-temporal en forma de pirámide disminuye el desempeño de las características propuestas en relación con el resultado presentado anteriormente. Este comportamiento podría ser atribuido a un *sobreajuste*[4] de la información presentada en entrenamiento debido al aumento en la dimensión del vector de características y el pequeño conjunto de entrenamiento.

1.4.2. Conjunto de video de Rochester

En un segundo experimento, se evaluó el algoritmo propuesto en el conjunto de videos de Rochester (Messing et al., 2009). Este conjunto de videos contiene 5 actores, desarrollando 10 acciones de la vida diaria que involucran 8 clases de objetos para un total de 150 videos. En comparación con el conjunto de videos

[3]Acrónimo tomado del inglés, "Average Precision"
[4]Traducido del término en inglés .ºverfitting"

Tabla 1.3: Evaluación cuantitativa de las características propuestas y de los esquema de representación propuesto y de referencia en el conjunto de videos de Rochester. Ver tabla 1.1 para más detalles.

Método	Precisión	AP promedio
Descriptor HOI Global (Prest et al., 2013)	90.7 %	93.9 %
Posición Relativa Global	82.0 %	92.7 %
Interacciones Globales	92.7 %	89.5 %
Posición Espacio-Temporal Relativa	92.7 %	90.1 %
Interacciones Espacio-Temporales	**98.0 %**	**94.2 %**

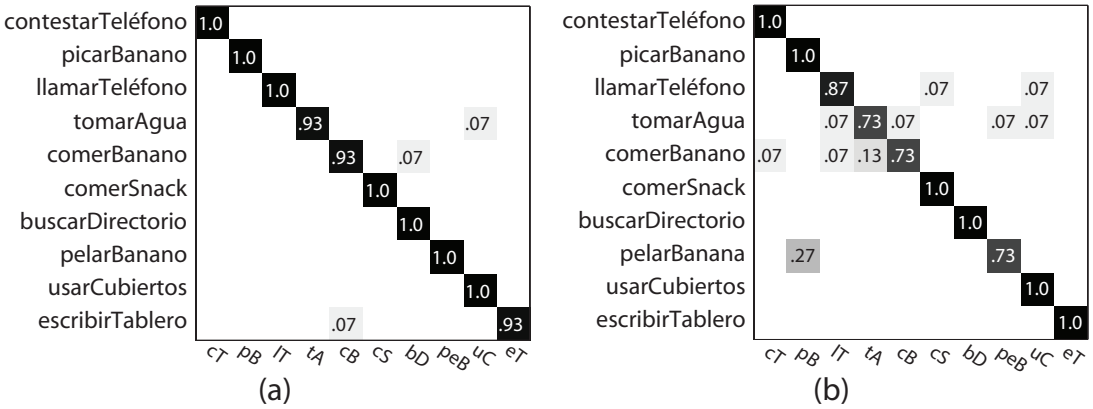


Figura 1.6: Matrices de confusión para el conjunto de videos de Rochester. (a) La descripción propuesta de las interacciones espacio temporales entre humano y objeto. (b) Descripción de la interacción humano-objeto propuesta por (Prest et al., 2013).

de gupta, este es un conjunto de videos más desafiante con videos grabados en una cocina real como escenario y objetos reales. En la figura 1.6(a) se resume el desempeño de la representación propuesta a través de una matriz de confusión. Igualmente, se compara el algoritmo propuesto con la implementación propia del algoritmo del estado del arte para la descripción de las interacciones humano-objeto (Prest et al., 2013), cuya matriz de confusión se aprecia en la figura 1.6(b). De manera similar a los resultados obtenidos anteriormente, se apreció que el descriptor espacio-temporal propuesto tiene un gran poder de discriminación en comparación con la mejor interacción reportada en la literatura (Prest et al., 2013). De igual forma, en la tabla 1.3 se aprecian las comparaciones de las características empleadas para el modelado de las interacciones y la representación de referencia de nuestro algoritmo.

Finalmente, la figura 1.6 muestra ejemplos de reconocimiento de acciones de manera exitosa, así como de errores de clasificación cometidos por el esquema

computacional propuesto. La mayoría de los errores se debe a una fuerte similitud de la relación espacio temporal entre humanos y objetos. En futuros trabajos se abordará esta situación con una descripción estructurada de la interacciones humano-objeto que tenga en cuenta la ocurrencia de aparición de múltiples objetos.

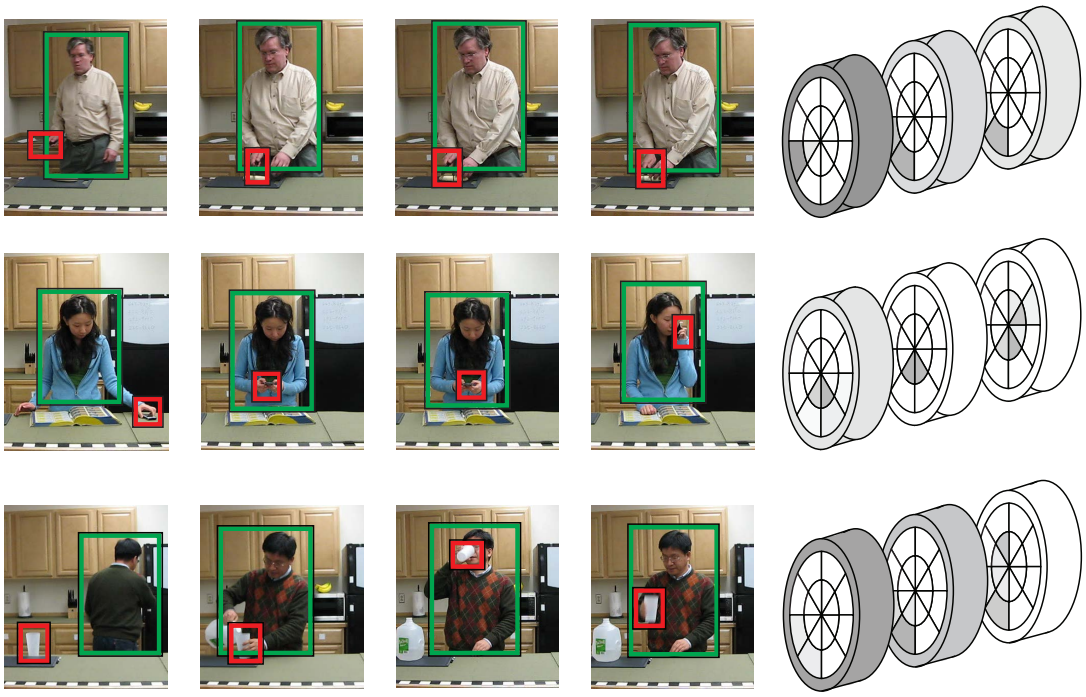


Figura 1.7: Resultados de ejemplos de reconocimiento. Las dos primeras filas muestran ejemplos de reconocimiento de acciones de manera exitosa en el conjunto de videos de Rochester. Se puede apreciar que que en ambos ejemplos, nuestra descripción de la interacción espacial-temporal entre humano y objeto captura correctamente la evolución de la interacción a través del tiempo. La última fila muestra un video clasificado de forma incorrecta por el el esquema de clasificación propuesto. En este caso, la descripción espacio-temporal de la interacción entre el actor y el vaso se confunde con la interacción de la acción asociada con usar el tenedor. En futuros trabajos se intentará ahondar en un descripción más estructurada del contexto de las interacciones entre humanos y objetos, con el fin de que el objeto que participe en la interacción tenga mayor relevancia sobre la decisión de clasificación.

Capítulo 2

Descomposición de las actividades humanas en términos de segmentos temporales discriminativos no solapadas

En el Capítulo 1 se demostró que el contexto temporal en el cual se desarrollan las interacciones espaciales entre humanos y objetos es ventajoso para mejorar la discriminación de las acciones humanas. Sin embargo, la forma en la que se definió la estructura temporal $\mathcal{V}$ de las acciones no es generalizable. De acuerdo con los resultados experimentales, ejemplificados en la figura 2.1 para la accion de *servir en una taza*, la alternativa de mejor desempeño es la de tres segmentos temporales uniformes no solapados. No obstante, ésto no implica que esta configuración sea óptima para otro conjunto de acciones. Además, el problema es más complejo si la duración de las acciones aumenta, *i.e.* análisis de actividades o eventos, puesto que la intuición de emplear segmentos uniformes no tendría mucha validez. Teniendo en cuenta estas vicisitudes y otras desventajas de la definición de estructuras temporales arbitrarias, este capítulo analiza un algoritmo que permita estudiar de manera particular la estructura temporal adecuada para una acción y en general de las actividades humanas. Si se desea establecer la estructura de las actividades humanas, un camino consecuente es la descomposición de las mismas en términos de segmentos temporales o de acciones atómicas. Éste es un tema que ha llamado la atención de varios investigadores en el campo de visión por computador (Pei et al., 2011),(Tang et al., 2012b),(Bobick and Davis, 2001),(Laxton et al., 2007), (Escorcia and Niebles, 2013),(Niebles et al., 2010). El interés en este tipo de representación está inspirado en la psicología de la Gestalt. De acuerdo con este punto

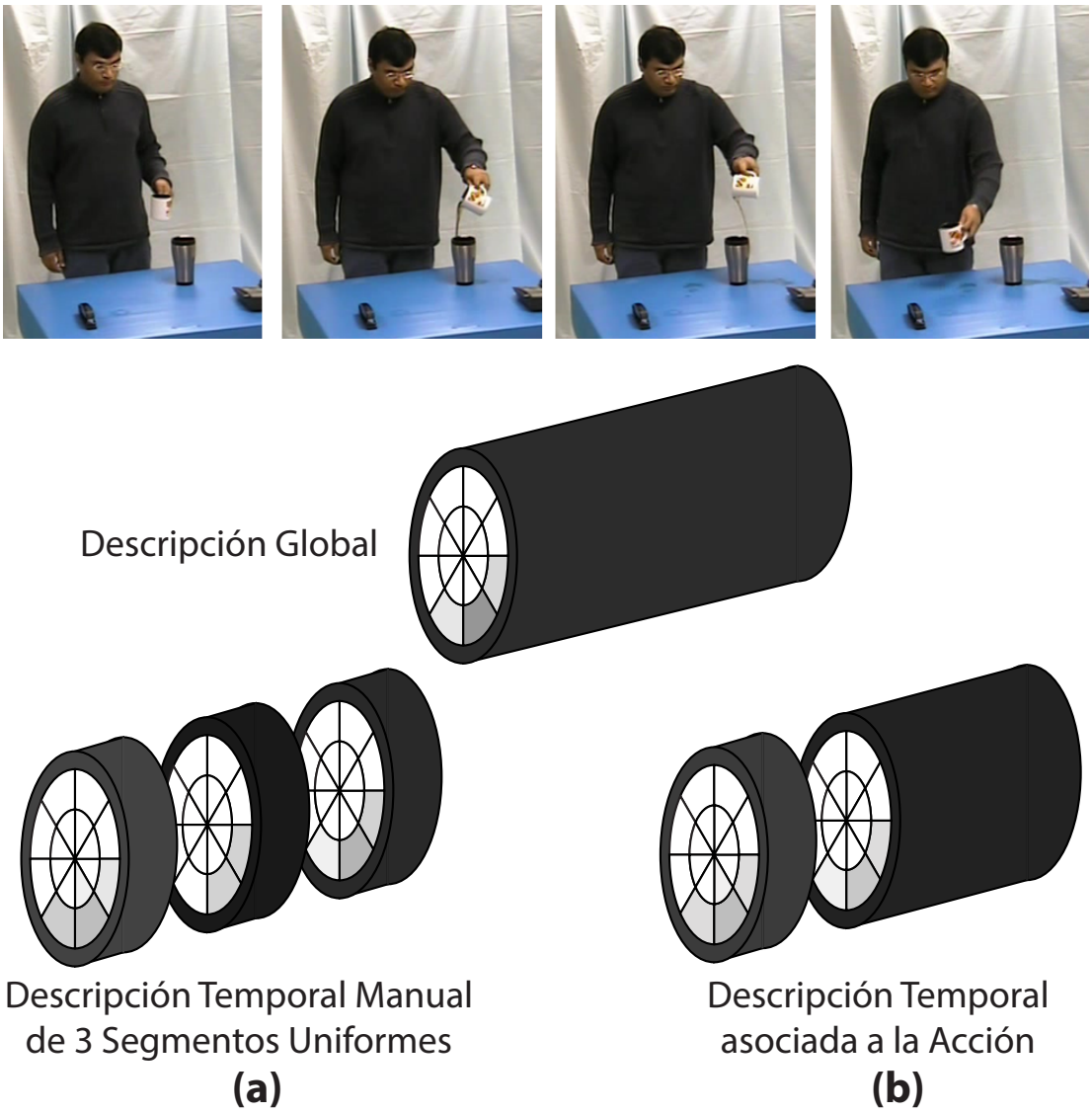


Figura 2.1: En el capítulo anterior se propuso una descomposición arbitraria de las acciones humanas de tal forma que se describa la evolución temporal de las interacciones (a). Sin embargo, no hay garantías de que este tipo de descripción sea óptima para todas las acciones. En este capítulo se analizará un algoritmo que permite capturar la estructura temporal discriminativa de las acciones humanas (b).

de vista, algunos psicólogos consideran la percepción de actividades como la extensión temporal de la percepción de objetos (Zacks and Tversky, 2001). Es decir, al analizar actividades como contestar el teléfono, el cerebro humano no solamente se enfoca en los actores que componen la acción, *i.e.* un teléfono y posiblemente una persona, también es importante el orden temporal en el que se desarrollan las sub-acciones, *i.e.* el repicar del teléfono, la persona buscando el teléfono, la acción de contestar la llamada y finalmente hablar por el teléfono. En algunos casos, si las sub-acciones no ocurren en el orden predeterminado para una actividad, el cerebro humano interpreta que ha sucedido otra actividad o que se encuentre frente a un evento anormal.

En la misma línea de las investigaciones psicológicas desarrolladas alrededor de la percepción de las actividades humanas presentadas en (Zacks and Tversky, 2001), en este capítulo se considera que la descripción del contexto temporal de las actividades es una pieza clave para el análisis y reconocimiento de las mismas. Más específicamente, se presentará un algoritmo que descompone de manera automática las actividades humanas en términos de segmentos temporales seleccionados de manera discriminativa. Cada segmento temporal se puede interpretar como una sub-acción de duración fija que compone la actividad y que ocurre en un intervalo de tiempo específico.

A diferencia de gran parte de los algoritmos actuales que modelan el contexto temporal de las actividades humanas, el algoritmo que se desea analizar decide de manera automática el número de segmentos temporales adecuados para la descomposición de una actividad. Esto representa dos ventajas frente a la mayoría de algoritmos actuales: (1) el algoritmo opera con un número de segmentos temporales adecuado para el reconocimiento de la actividad, (2) cada actividad se descompone de manera específica en relación con sus características dinámicas frente a las de las demás actividades. La desventaja de trabajar con un número de segmentos temporales o sub-acciones erróneo es que el desempeño de clasificación del algoritmo puede verse comprometido, puesto que en la mayoría de estos algoritmos, todos los segmentos temporales definidos deben ocurrir para que la actividad sea detectada. Por lo tanto, al definir un gran número de segmentos temporales es posible causar un *sobreajuste* de las observaciones durante entrenamiento. Mientras que, definir un número pequeño de segmentos temporales puede ocasionar que el algoritmo no tenga la capacidad de discriminación suficiente para distinguir las actividades. Por otro lado, en la mayor parte de los trabajos todas las categorías de actividades se descomponen en término del mismo número de segmentos temporales. Esto obliga a que el algoritmo se enfoque en describir el orden temporal de las segmentos temporales, en lugar de describir de forma compacta la estructura dinámica de la actividad.

A pesar de las desventajas de usar un número inadecuado de segmentos temporales, definir el número de segmentos a través de los cuales se representará una actividad no es tan trivial como parece, puesto que el conocimiento del dominio no garantiza un buena decisión. Es cierto que, un mayor número de segmentos temporales permite capturar la evolución temporal en un nivel más fino. Sin embargo, es la información capturada durante los intervalos asociados al segmento temporal la que determina la capacidad de discriminación del modelo. Por tal motivo, una definición adecuada del número de segmentos temporales no solo depende de la naturaleza del dominio, también depende de la naturaleza de los descriptores empleados. A diferencia del diseñador del algoritmo, cuya asignación del número de segmentos temporales esta sesgado por el conocimiento de la actividad, el algoritmo propuesto escoge la manera en la que se descomponen las actividades de manera automática. Para ello, se realiza una selección discriminativa de los segmentos temporales basada en los descriptores visuales utilizados para la caracterización de los mismos. Esta metodología de selección concuerda con la tendencia de los trabajos recientes enfocados en la construcción de partes discriminativas para el reconocimiento de objetos (Bourdev et al., 2010),(Singh et al., 2012). Al igual que estos trabajos, el algoritmo propuesto reconoce que existe una gran brecha entre las características visuales y la información semántica relacionada con el contenido de la acción. Por tal motivo, las partes semánticas de los objetos o las actividades no necesariamente son capaces de aumentar la capacidad de discriminación, debido a que estas se describen en términos las características visuales. En ese sentido, en lugar de guiarse por como concebimos que esta compuesta una actividad, los trabajos recientes sugieren que resulta más benéfico encontrar las sub-acciones de la actividad basados en su poder discriminativo.

Este capítulo esta organizado de la siguiente manera: en la sección 2.1 se discutirá el trabajo previo relacionado con el análisis y clasificación de las actividades con base en el contexto temporal de las mismas. Luego, en la sección 2.2 se presentará el algoritmo que permite la descomposición automática de las actividades humanas en término de segmentos temporales no solapados. Posteriormente, en la seccion 2.3 se describe como se integra el algoritmo presentado en la seccion 2.2 dentro del esquema de aprendizaje de maquina. Finalmente, en la sección 2.4 se analizarán las evidencias experimentales obtenidas con este algoritmo.

2.1. Trabajo previo

En esta sección se presentarán varios trabajos relacionados con el análisis y reconocimiento de actividades enfocados en la descripción de la evolución temporal

de las mismas. Con el fin de facilitar su descripción y comparación, los trabajos revisados han sido agrupados en cuatro amplias categorías.

2.1.1. Descomposición de acciones y eventos de forma no supervisada

El uso de técnicas no supervisadas que permiten descomponer las actividades o eventos en términos de acciones simples o subconjuntos de segmentos temporales de menor duración ha captado la atención de muchos investigadores en el área de visión por computador y síntesis de gráficos y videos (Vecchio et al., 2003),(Barbic et al., 2004).

Zhou *et. al* propusieron un algoritmo jerárquico para la segmentación de una secuencia de movimientos humanos basado en técnicas de agrupamiento sobre múltiples segmentos dentro de la secuencia (Zhou et al., 2013). De esta forma, una secuencia de movimientos humanos tales como *caminar-saltar-caminar* se puede dividir en tres segmentos temporales enfocados en los movimientos atómicos simples. La naturaleza jerárquica del algoritmo permite que cada segmento atómico sea dividido en segmentos temporales más cortos, los cuales podrían carecer de algún significado semántico para la acción de interés. De forma similar, Nater *et. al* propusieron un algoritmo para la segmentación de secuencias de videos de larga duración a través de un algoritmo jerárquico de agrupamiento basado en *análisis de características suaves*, SFA[1], (Nater et al., 2011). Este trabajo se enfoca en la detección de eventos inusuales en nuevas secuencias de video a partir de la síntesis de la secuencia de interés con la estructura de árbol aprendida de una secuencia de video previa. Por otro lado, Pei *et. al* proponen un método no supervisado para aprender la estructura de eventos de larga duración en termino de acciones o interacciones atómicas conocidas (Pei et al., 2011),(Si et al., 2011). El algoritmo de Pei *et. al* hace uso de una *gramática probabilística libre de contexto*, SCFG[2], representada a través de grafos AND-OR temporales, esto les permite sintetizar la información de contexto temporal en diferentes escalas y compensar algunos errores de detección en las acciones atómicas. Al igual que el método de (Nater et al., 2011), a partir de la estructura aprendida se pueden interpretar los eventos en una nueva secuencia de video.

A diferencia del algoritmo que se desea analizar para la descomposición de las actividades, estos modelos construyen la estructura de un evento con base en único video. Es decir, su interés no es describir la estructura asociada a una actividad

[1] Acrónimo tomado del inglés "Slow Feature Analysis"
[2] Acrónimo tomado del inglés "Stochastic Context-Free Grammar"

en particular. Estos trabajos intentan: (1) descomponer un video en secuencias más simples o (2) analizar un video en términos de la una estructura aprendida previamente. De esta forma, estos algoritmo pueden ser de gran utilidad para la síntesis de movimiento de juegos de vídeo y el análisis de los eventos anormales en una estación de vigilancia. Sin embargo, su aplicación no es directa en el área de categorización y modelado de actividades.

2.1.2. Modelos temporales basados en cadenas de Markov

Al discutir sobre el modelado temporal de las acciones es imprescindible mencionar los trabajos basados en la teoría de cadenas de Markov, apetecidos dentro de la comunidad científica gracias a su exitosa aplicación para la tarea de reconocimiento de voz. Los primeros trabajos hicieron uso de modelos basados en máquinas de estado finitos o métodos de aprendizaje generativos con estados latentes (Wilson and Bobick, 1999). Por otro lado, trabajos más recientes han explorado métodos de aprendizaje discriminativos basados en CRF (Wang et al., 2006). Asimismo, algunos investigadores han usado modelos semimarkovianos que permiten introducir dependencias temporales entre acciones (Nguyen et al., 2011) o dependencias composicionales entre sub-acciones que pertenecen a una acción (Tang et al., 2012a).

Wang *et. al* propusieron un modelo temporal para el reconocimiento de gestos humanos denominado *campos aleatorios condicionales ocultos*, HCRF[3], (Wang et al., 2006). Este modelo es una versión discriminativa de las *Cadenas ocultas de Markov*, HMM[4], con mejor eficacia para la tarea de reconocimiento. Recientemente, Tang *et. al* propusieron un modelo HCRF entrenado con técnicas de maximización del margen en el cual cada actividad se representa a través de un conjunto de sub-acciones latentes, las cuales, a su vez poseen un conjunto de duraciones latentes (Tang et al., 2012a). De esta forma, el modelo para el reconocimiento de las actividades es flexible en cuanto a la transición entre sub-acciones latentes y en cuanto a la duración de las mismas.

A diferencia del algoritmo de descomposición de actividades que se analizará en este capítulo, los algoritmos basados en cadenas de Markov deben especificar el número de sub-acciones latentes que componen una actividad o la representación visual asociada con las sub-acciones. En cualquiera de los dos casos, estos modelos sufren las desventajas mencionadas anteriormente con respecto a la asignación

[3]Acrónimo tomado del inglés “Hidden Conditional Random Fields”
[4]Acrónimo tomado del inglés “Hidden Markov Model”

de un número de segmentos temporales inadecuados. Por lo tanto, en un trabajo futuro se puede explorar cómo reformular este tipo de modelos, de tal forma que el número de sub-acciones latentes sea aprendido de manera discriminativa.

2.1.3. Modelos jerárquicos

A diferencia de los modelos basados en cadenas de Markov, donde cada actividad se representa como una secuencia de acciones, algunos investigadores han propuestos modelos composicionales más profundos. Estos modelos permiten que las actividades se representen como una composición acciones más simples (Moore and Essa, 2002), algunos de ellos están inspirados en los modelos de sintaxis gramatical empleados en el procesamiento del lenguaje (Jurafsky and Martin, 2008). Ivanov y Bobick hacen uso de un SCFG para el reconocimiento de acciones humanas complejas en términos de acciones simples (Ivanov and Bobick, 2000). La tarea del algoritmo SCFG, es analizar el conjunto de acciones simples detectadas en un video a través de un gran número de reglas probabilísticas de producción y escoger la acción compleja que mejor explica el conjunto de símbolos terminales detectados.

La desventaja de los modelos basados en SCFG es que la noción de orden temporal es muy débil. Por tal motivo, algunos investigadores hacen uso de gramáticas de contexto sensibles aumentando la complejidad del algoritmo de inferencia (Tran and Davis, 2008). Por otro lado, existe otro tipo de modelos estructurales que permiten el análisis de secuencias temporales en diferentes niveles de abstracción. Laxton *et. al* proponen el uso de un *red bayesiana dinámica*, DNB[5], para codificar de manera parcial el orden entre sub-acciones y que permite el reconocimiento de acciones complejas (Laxton et al., 2007).

Recientemente, Brendel y Todorovic construyeron grafos espacio-temporales de videos que codifican relaciones jerárquicas, espaciales y temporales de las actividades humanas (Brendel and Todorovic, 2011). Por su parte, Gaidon *et. al* presentaron un algoritmo jerárquico que resume la información visual de un video en términos de un árbol temporal (Gaidon et al., 2012), cada hoja del árbol representa un conjunto de trayectorias densas (Wang et al., 2011). A diferencia de los trabajos anteriores que hacen uso de acciones primitivas o agentes semánticos como humanos y objetos, (Brendel and Todorovic, 2011) y (Gaidon et al., 2012) trabajan con bloques visuales de bajo nivel. Por lo tanto, luego de la etapa de entrenamiento no es posible obtener una estructura jerárquica canónica de la actividad.

[5]Acrónimo tomado del inglés “Dynamic Bayesian Network”

2. Descomposición de las actividades humanas en términos de segmentos temporales discriminativos no solapadas

A diferencia de los modelos basados en grámaticas y redes bayesianas, el algoritmo que se desea estudiar para la descomposición no requiere la definición de acciones atómicas para la construcción de la estructura temporal, puesto que los segmentos temporales que componen una actividad son determinadas de manera discriminativa. Por otro lado, los trabajos recientes que emplean grafos temporales asocian un árbol de segmentación a cada video y determinan la presencia de una actividad, en la medida en que el árbol de segmentación presente características estructurales y visuales similares a las de los arboles de los videos de entrenamiento. Es decir, son modelos inductivos, mientras que, el algoritmo que se estudiará es de carácter deductivo, dado que construye la estructura de la actividad a partir de las observaciones o segmentos temporales comunes en todos los videos asociados con una actividad particular.

2.1.4. Modelos de segmentos temporales y con estructuras tipo árbol

Gran parte de los trabajos enfocados en el reconocimiento de acciones simples en videos con pocas restricciones hacen uso del modelo de *bolsa de palabras*, BOW, con diferentes tipos de descriptores visuales (Niebles et al., 2008), (Wang et al., 2009). Este tipo de modelo acumula de forma global la aparición de los bloques visuales representativos presentes en cada video sin tener en cuenta ninguna relación de orden temporal. Es decir, este tipo de modelo considera que a cada acción le corresponde un único segmento temporal indivisible. Inspirados en las descripciones espaciales de las escenas en imágenes (Lazebnik et al., 2006), algunos investigadores complementan la descripción global con partes espacio-temporales rígidas integradas a través de *kernels* (Laptev et al., 2008).

Por otro lado, algunos autores inspirados en el modelo de partes discriminativas deformables para el reconocimiento de objetos (Felzenszwalb et al., 2010) propusieron descomponer las acciones y actividades humanas en términos de p segmentos temporales flexibles. Gaidon *et. al* propusieron un algoritmo para el reconocimiento de acciones sencillas simples a través de la detección de p segmentos temporales atómicos con un orden temporal específico (Gaidon et al., 2011). A diferencia de las partes temporales empleadas por (Escorcia and Niebles, 2013), cada segmento temporal es flexible en cuanto a duración y localización. De manera similar, Niebles *et. al* propone la descomposición de actividades en términos de segmentos flexibles seleccionados de manera discriminativa a través de un algoritmo *latente con máquinas de soporte vectorial*, LSVM[6], (Niebles et al., 2010).

[6]Acrónimo tomado del inglés "Latent Support Vector Machine"

A diferencia del algoritmo de (Gaidon et al., 2011), esta propuesta: (1) selecciona de forma automática los p segmentos temporales en los cuales se descomponen las acciones sin necesidad de anotaciones humanas adicionales; (2) el modelo considera la contribución del segmento global donde se enmarca la acción, al igual que la contribución de cada una de las partes, desembocando en una estructura tipo árbol de profundidad uno; (3) el uso de un algoritmo discriminativo para la selección automática de las partes no garantiza que las mismas tengan un significado semántico como en los segmentos temporales de (Gaidon et al., 2011).

Recientemente, Ryoo y Matthies propusieron un algoritmo jerárquico que permite el aprendizaje de la estructura de las actividades de manera discriminativa (Ryoo and Matthies, 2013). A diferencia de los trabajos enunciados anteriormente, este algoritmo determina de manera automática la estructura de las actividades y es capaz de definir en cuantos segmentos temporales no solapados se puede descomponer la misma. Este algoritmo se introdujo en el contexto de reconocimiento de acciones egocéntricas. Las cuales se describen a través de descriptores aditivos que codifican patrones de movimiento globales y de puntos de interés en la escena. Nuestro interés es extender su uso al aprendizaje de las estructuras espacio-temporales de las interacciones entre humanos y objetos. Asimismo, se discutirán algunas modificaciones que permiten mejorar su rendimiento.

2.2. Descomposición de actividades humanas en término de segmentos temporales discriminativos

De acuerdo con los resultados presentados en el capítulo 1 y los trabajos previos presentados en la sección 2.1, no cabe duda que modelar el contexto temporal de las actividades humanas es crucial para obtener una categorización adecuada de actividades similares. Sin embargo, gran parte de los trabajos actuales que modelan la evolución temporal de las actividades, definen el número de segmentos temporales que componen la actividad como un parámetro externo, el cual debe ser definido por el diseñador del algoritmo.

Teniendo en cuenta que las dificultades para la definición del número de segmentos temporales pueden ir más allá del conocimiento del dominio del diseñador del algoritmo, resulta ideal emplear un algoritmo de aprendizaje que determine de manera automática el numero de segmentos temporales a través de las cuales se puede descomponer una actividad.

2. Descomposición de las actividades humanas en términos de segmentos temporales discriminativos no solapadas

En esta sección se presentará una modificación del algoritmo de Ryoo y Matthies (Ryoo and Matthies, 2013) que permite la descomposición iterativa de las actividades en términos de segmentos temporales discriminativos no solapadas, de tal forma que se obtiene una mejor discriminación de las mismas. El resto de la sección esta organizada de la siguiente forma: (1) primero se describirá la forma en la cual se representan los videos de las actividades humanas; (2) posteriormente se presentará la estructura de las actividades humanas y la función de *kernel* que captura la similitud entre dos videos; (3) luego se describirá el algoritmo que permite aprender la estructura de las actividades humanas; y (4) por último se citan de manera explícita las modificaciones realizadas con respecto al algoritmo original propuesto en (Ryoo and Matthies, 2013).

2.2.1. Descripción de los videos

El algoritmo empleado para determinar la estructura de las actividades puede ser aplicado sobre una gran variedad de descriptores visuales. En general, la única restricción es que el descriptor utilizado para la representación de una secuencia pueda ser computado en múltiples escalas temporales. Las representaciones calculadas al nivel de los frames del video o que desembocan en histogramas de ocurrencia son ejemplos de descriptores factibles con los que puede operar el algoritmo.

Sin perder generalidad, de aquí en adelante se asumirá que un video i de duración T_i se puede representar a través de un conjunto secuencial de T_i características de dimensión d. De esta forma, se define x_i como el descriptor visual asociado con el video i de la siguiente forma $x_i = \left\langle x_i^1, x_i^2, \ldots, x_i^t, \ldots, x_i^{T_i} \right\rangle | x_i^t \in \mathbb{R}^d$.

2.2.2. Estructura de las actividades y *kernel* de alineamiento

Una actividad se representa por medio de la concatenación de sus segmentos temporales. De esta forma, la estructura de una actividad S consiste en un conjunto de divisiones particulares que dividen la duración de toda la secuencia de video en término de múltiples segmentos de video.

El objetivo de la representación estructural de las actividades es explotar el contexto temporal de los segmentos temporales en una actividad de tal forma que al contrastar los segmentos temporales en el orden dispuesto por la estructura S dos videos que contienen la misma actividad su similitud sea mayor a la de un par

de videos con actividades diferentes. En la figura 2.2 se esquematiza la estructura de una secuencia sintética de componentes y como a través del orden temporal impuesto por los segmentos temporales se puede comparar dos secuencias sintéticas. Es interesante apreciar que la estructura utilizada en esta figura, concluye que las secuencias son muy similares. Por tal motivo, es importante realizar una definición adecuada de la estructura de las actividades. En la figura 2.3 se presenta la ventaja de realizar una adecuada definición de la estructura temporal para distinguir secuencias sintéticas.

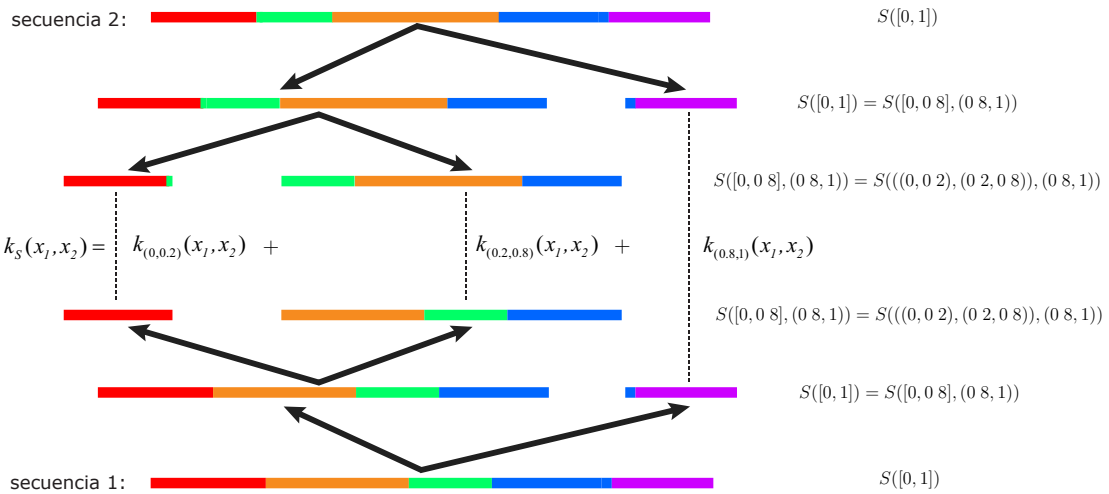


Figura 2.2: A partir de la estructura de una actividad $S\left(((0,0{,}2),(0{,}2,0{,}8)),(0{,}8,1)\right)$ es posible comparar las secuencias 1 y 4 obedeciendo al orden temporal de los segmentos temporales, segmentos terminales encerrados entre paréntesis, que componen la actividad.

De manera formal, se puede representar la estructura de una actividad en términos de divisiones jerárquicas binarias con las siguientes reglas de producción:

$$\begin{array}{lcl} [t_1,t_2] & \rightarrow & ([t_1,t_3],[t_3,t_2]) \\ [t_1,t_2] & \rightarrow & (t_1,t_2) \end{array} \tag{2.1}$$

donde t_3 es un punto de tiempo relativo $(0 \leq t_1 < t_3 < t_2 \leq 1)$ que describe como la estructura de la actividad divide el video en el intervalo de tiempo comprendido entre $[t_1,t_2]$. Cada intervalo de tiempo (t_1,t_2) generado a partir de la segunda regla de producción se denomina segmento terminal, especificando que la estructura de la actividad lo considera una sub-acción. Es decir, es un segmento indivisible a través del cual se describe la actividad.

Cualquier estructura de una actividad construida aplicando de manera sucesiva las reglas de producción de la ecuación 2.1 iniciando en $S\,[0,1]$ hasta cubrirlo únicamente por segmentos terminales, se considera una estructura válida. Ejemplos de estructuras válidas se aprecian en las figuras 2.2 y 2.3, mientras que una estructura inválida puede adoptar la siguiente forma $S^{'}([0,0{,}5],(0{,}5,1))$. La estructura $S^{'}$ se considera inválida, debido a que esta compuesta por un segmento terminal ubicado en el intervalo $(0{,}5,1)$ y un segmento no-terminal correspondiente al intervalo $[0,0{,}5]$. Para que dicha estructura se considere válida, es necesario que se sigan aplican las reglas de la ecuación 2.1. De tal manera que el segmento no terminal se exprese en término de otros segmentos temporales.

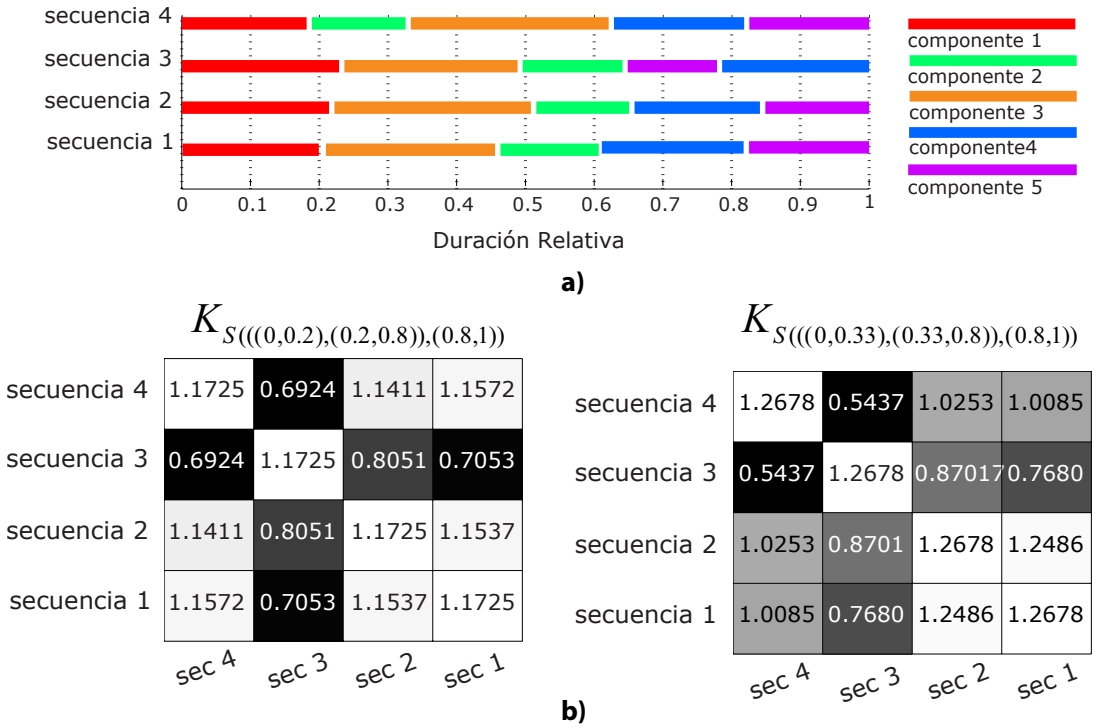


Figura 2.3: Comparación de secuencias sintéticas a través de la estructura de una actividad. (a) Dado el conjunto de secuencias sintéticas, es factible que al definir una estructura adecuada para una categoría, las secuencias con una dinámica temporal similar adquieran una similitud más alta en relación a secuencias con un dinámica distinta. (b) Nótese que, con la estructura de actividad $S\left(((0,0{,}33),(0{,}33,0{,}8)),(0{,}8,1)\right)$ las secuencias 1 y 2, cuyas componentes aparecen en el mismo orden y con similar duración, tienen una similitud alta entre sí en relación con las otras secuencias. Mientras que, con la estructura de actividad $S\left(((0,0{,}2),(0{,}2,0{,}8)),(0{,}8,1)\right)$ la secuencia 4 también tiene una dinámica temporal similar a la de las secuencias 1 y 2.

Dada una estructura de actividad S, el *kernel* asociado con ésta $k_S\,(x_i,x_j)$

mide la similitud entre los descriptores visuales asociados a los videos x_i y x_j a partir del siguiente par de ecuaciones:

$$\begin{aligned} k_{[t_1,t_3],[t_3,t_2]}(x_i, x_j) &= k_{[t_1,t_3]}(x_i, x_j) + k_{[t_3,t_2]}(x_i, x_j) \\ k_{(t_1,t_2)}(x_i, x_j) &= k(x_i^{(t_1,t_2)}, x_j^{(t_1,t_2)}) \end{aligned} \tag{2.2}$$

donde $x_i^{(t_1,t_2)}$ representa el descriptor visual del video i computado en el intervalo de tiempo (t_1, t_2) y $k(x, y)$ es la función de kernel que calcula la similitud entre los descriptores visuales x y y, y satisface el teorema de Mercer (Burges, 1998).

La ecuación 2.2 implica que la similitud entre dos videos es equivalente a la suma de las similitudes de entre los segmentos terminales o nodos terminales que componen la actividad.

2.2.3. Aprendizaje de la estructura de una actividad

En esta sección se presenta el algoritmo para el aprendizaje de la estructura de una actividad S a partir de m videos de entrenamiento. El resto de la subsección esta organizado de la siguiente forma. En primer lugar, se definirá la noción de *alineamiento de kernels* (Cristianini et al., 2001), (Cortes et al., 2012) a partir de la cual se evaluará el poder discriminativo del kernel asociado con la estructura de la actividad S para la tarea de reconocimiento de actividades. La idea es constrastar el kernel asociado con una estructura de actividad contra una "función de kernel óptima". Posteriormente, se presentará el algoritmo que permite el aprendizaje de la estructura óptima de una actividad, el cual evalúa de forma jerárquica múltiples estructuras candidatas haciendo uso del alineamiento de kernels.

Alineamiento de kernels: a partir de un conjunto de datos de entrenamiento $(x_1, x_2, ..., x_m)$ es posible obtener la matriz de Gram K para un función de kernel k como:

$$K = \left(k\left(x_i, x_j\right)\right)_{i,j=1}^{m} \tag{2.3}$$

El *alineamiento* entre dos funciones de kernel k_1 y k_2 mide que tan similares son los valores de similitud calculados por cada función de kernel. De manera rigurosa, esta se define como:

$$A\left(K_1, K_2\right) = \frac{\langle K_1, K_2 \rangle_F}{\|K_1\|_F \|K_2\|_F} \tag{2.4}$$

donde $\langle K_1, K_2 \rangle_F$ es el producto interno Frobenius entre las matrices de kernel k_1 y k_2. Lo que es equivalente a $\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^{m} k_1(x_i, x_j) k_2(x_i, x_j)$. Por otro lado, $\|K_1\|_F$ corresponde a la norma Frobenius de la matrix K_1, la cual es equivalente a la raíz cuadrada del producto Frobenius de la matrix con ella misma.

A partir de la noción de *alineamiento de kernels* es posible evaluar diferentes estructuras para una actividad. Con este propósito en mente, se definirá una "función de kernel óptima"l con respecto a la cual se medirá el alineamiento de cualquier estructura particular S. De esta forma, se define una métrica a partir de la cual se puede escoger la estructura S^* del conjunto $\mathcal{S}$ de posibles estructuras candidatas para una actividad.

Nuestra *función de kernel óptima* l esta definida de tal forma que dos videos asociados con la misma actividad poseen similitud alta, mientras que dos videos cuyas actividades son diferentes poseen una similitud baja. A continuación se presenta de manera formal nuestra función de kernel óptima y su matriz de Gram L

$$\begin{array}{lcl} l(i,j) & = & \begin{cases} 1 & y_i = y_j \\ -1 & \text{otro caso} \end{cases} \\ L & = & \left(l\left(i,j\right)\right)_{i,j=1}^{m} \\ y_i & \in & \{1,-1\} \end{array} \tag{2.5}$$

La etiqueta $y_i = 1$ representa que en el i-ésimo video de entrenamiento se desarrolla la actividad de interés a. En tanto que, la etiqueta $y_i = -1$ indica que durante el i-ésimo video de entrenamiento se desarrolla una actividad no relevante.

Al computar el *alineamiento de kernels* entre el kernel asociado con la estructura de actividad S y nuestra matrix de kernel óptimo L i.e. $A(K_S, L)$, se mide el grado de similitud del kernel K_s en relación con el kernel óptimo. De manera intuitiva, se puede apreciar que una medida de alineamiento alto indica que la estructura S asociada con la actividad favorece la adecuada clasificación de los videos de entrenamiento. Por tal motivo, se considera que la medida de alineamiento $A(K_S, L)$ es una función de selección discriminativa.

La noción de *alineamiento de kernel* de la ecuación 2.4 es una medida que resume las observaciones de manera uniforme. Este tipo de medidas resultan inadecuadas para gran parte de los algoritmos de reconocimiento de patrones, donde el número de observaciones irrelevantes suele ser considerablemente mayor al número de observaciones de interés, puesto que el algoritmo de reconocimiento encausa su aprendizaje en no cometer errores con las observaciones negativas. Teniendo esto en mente, es necesario reformular la definición del *alineamiento de kernels* para sostener el desbalanceo inherente en la recolección de los videos (Cortes et al., 2012). Para ello, en lugar de computar el alineamiento directamente con las matrices de

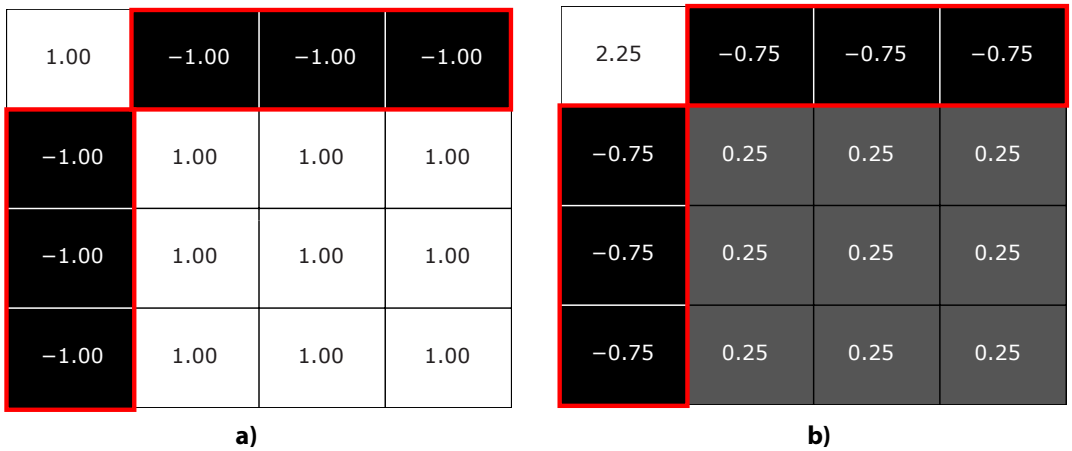


Figura 2.4: Ilustración de la propiedad del centrado sobre la matriz de kernel óptima para un conjunto de etiquetas $y = [1,\ -1,\ -1,\ -1]^T$. De izquierda a derecha se aprecian la matriz de kernel óptima original (a) y la matriz de kernel óptima centrada (b). Como se puede apreciar al sumar las contribuciones en la "zona discriminativa"de la matriz de kernel sin centrar, resaltada en rojo, se observa que son inferiores a las de la "zona de alta similitud", zonas no resaltadas. Mientras que, en la matriz de kernel centrada la contribuciones de ambas zonas son equitativas. Dado que en general, el número de instancias no relevantes es mayor al número de instancias de interés, una matriz de kernel sin centrar prefiere estructuras donde el conjunto de datos negativos sea uniforme.

Gram K_1, K_2 se emplean las matrices de Gram centradas K_{c1}, K_{c2}.

$$\begin{aligned} A\left(K_{c1}, K_{c2}\right) &= \frac{\langle K_{c1}, K_{c2}\rangle_F}{\|K_{c1}\|_F \|K_{c2}\|_F} \\ K_c &= CKC \\ C &= I_m - \frac{1}{m} 1 \times 1^T \end{aligned} \tag{2.6}$$

donde I_m es la matriz idéntica de tamaño m y 1 representa un vector columna de tamaño m. En la figura 2.4 se aprecia la propiedad de centrado de la matriz de Gram sobre la función de kernel óptima.

Sin perder generalidad y con el ánimo de facilitar la comprehensión y descripción del resto de la subsección se denotará $A(K_S, L)$ simplemente como $A(K_S)$.

Aprendizaje jerárquico de la estructura de una actividad: a continuación se presenta el algoritmo de búsqueda de la estructura óptima de una actividad con base en el conjunto de videos de entrenamiento. El objetivo del algoritmo es retornar la estructura óptima S^* para una actividad a que maximiza el alineamiento de kernels para un conjunto de videos de entrenamiento $s^* = \underset{S}{\operatorname{argmax}} A(K_S)$.

Mas específicamente, el algoritmo de aprendizaje de la estructura óptima de

una actividad $S[0,1]^*$ consiste en encontrar la mejor estructura que divida el intervalo de duración completo de la actividad $[0,1]$ y se encuentra resumido en la siguiente ecuación:

$$S[t_1,t_2]^* = \underset{S[t_1,t_2]}{\operatorname{argmax}} \left\{ \max_{\substack{t \\ t_1<t<t_2}} A\left(K_{(S[t_1,t]^*,S[t,t_2]^*)}\right), A\left(K_{S(t_1,t_2)}\right) \right\} \tag{2.7}$$

Nótese que, el conjunto de estructura candidatas que se desprende de la ecuación 2.7 es extremadamente grande, puesto que para determinar el óptimo de la maximización interna es necesario determinar las particiones óptimas de los intervalos $[t_1,t]$, $[t,t_2]$. Los cuales a su vez requieren resolver un nuevos problemas de optimización.

En lugar de explorar el conjunto exponencial de estructuras candidatas, es posible emplear técnicas de optimización como A^* o de *ramificación y poda* para encontrar el óptimo de la maximización interna en la ecuación 2.7. Por otro lado, algunos trabajos han demostrado que las *aproximaciones voraces* son igualmente efectivas y más eficientes en algunos casos (Desai et al., 2009). Por tal motivo, se propone realizar la siguiente aproximación voraz para encontrar el óptimo de la maximización interna:

$$\underset{t}{\operatorname{argmax}}\, A\left(K_{(S[t_1,t]^*,S[t,t_2]^*)}\right) \approx \underset{t}{\operatorname{argmax}}\, A\left(K_{(S(t_1,t),S(t,t_2))}\right) \tag{2.8}$$

De esta forma, al explorar el intervalo $[t_1,t_2]$ se asume que las dos divisiones que se desprendan del mismo serán segmentos terminales de la estructura. De la ecuación 2.8 se desprende la siguiente ecuación recursiva T que permite obtener la estructura óptima S^* para una actividad a partir de $T[0,1]$:

$$\begin{aligned} T[t_1,t_2] &= \begin{cases} (T[t_1,t_3],T[t_3,t_2]) & t_3 \neq t_1,t_2 \wedge (\mathrm{Err}_d < \mathrm{Err}_o) \\ (t_1,t_2) & \text{otro caso} \end{cases} \\ t_3 &= \underset{t}{\operatorname{argmax}}\, A\left(K_{(S(t_1,t),S(t,t_2))}\right) \end{aligned} \tag{2.9}$$

donde **Err** es el porcentaje de error de reconocimiento de la estructura temporal de la actividad computada usando los datos de entrenamiento. $\mathbf{Err_d}$ es el porcentaje de error de reconocimiento de la estructura temporal de una actividad, en la cual se acaba de introducir una división en el intervalo $[t_1,t_2]$. $\mathbf{Err_o}$ corresponde al error de reconocimiento de la estructura temporal de una actividad sin introducir ninguna partición en el intervalo (t_1,t_2).

En resumen, el algoritmo propuesto para el aprendizaje de la estructura de una actividad evalúa de manera independiente el particionamiento de todos los

segmentos no terminales en una estructura. Asumiendo que: (1) al particionar un segmento se originan dos segmentos terminales, (2) el resto de segmentos no terminales en la estructura son segmentos terminales. De todas las posibles particiones evaluadas, incluyendo la estructura actual de la actividad sin introducir partición, se selecciona aquella cuyo kernel este mejor alineado con la función de kernel óptimo siempre y cuando su porcentaje de error de reconocimiento sea menor a la de la estructura actual sin introducir particiones.

Por último, en algunos casos es posible que aún con el centrado de las matrices de kernel, el algoritmo propuesto descarte particiones discriminativas en niveles más profundos de la estructura. Por tal motivo, es válido plantear una versión más competitiva del algoritmo concebido hasta ahora, tal que durante cada partición se enfoque únicamente en los videos alineados incorrectamente. Para ello, (1) en lugar de computar el alineamiento de toda la matriz de kernel, éste se calcula sobre la porción discriminativa de las matrices de kernel[7]; y (2) las etiquetas de los videos no relevantes alineados correctamente, se eliminan de la matriz de kernel óptimo. Vale la pena anotar que esta versión del algoritmo es más sensible a causar *sobre-ajuste* en los datos. Sin embargo, como se apreciará en la sección experimental es posible obtener resultados satisfactorios con esta versión más competitiva. De aquí en adelante, se asumirá que el algoritmo propuesto incluye estas últimas modificaciones a menos que se especifique lo contrario.

A manera de resumen y guía para el programador se presenta el pseudocódigo del algoritmo de aprendizaje de la estructura de una actividad en el algoritmo 1. Nótese que la inicialización de la estructura de la actividad consiste en crear un único nodo que abarque el segmento $[0, 1]$. Asimismo, es posible disminuir el tiempo de los procesos de búsqueda y listado considerando que los nodos ramificados son candidatos a ser explorados y nodos no-terminales.

2.2.4. Modificaciones realizadas al algoritmo de aprendizaje de la estructura de las actividades

A continuación se citan de manera explícita las modificaciones realizadas al algoritmo de aprendizaje propuesto por Ryoo y Matthies (Ryoo and Matthies, 2013).

1. Función de kernel óptimo: a diferencia de la función de kernel óptimo de Ryoo y Matthies, nuestra función de kernel óptimo tiene en cuenta la contribución

[7]La porción discriminativa de la matriz de kernel es aquella que relaciona los videos de la categoría de interés con los videos de las otras categorias

de todas las zonas de la matriz de kernel, asignando similitudes altas o bajas, de acuerdo con la zona específica.

2. Se definió la noción de *alineamiento de kernels* en término de las matrices de kernel centradas: una desventaja del algoritmo propuesto en (Ryoo and Matthies, 2013) es que el aprendizaje de la estructura de una actividad se debe realizar de tal forma que exista un número igual de videos con etiquetas positivas y negativas. En la práctica, para no afectar la capacidad de generalización del algoritmo, esta restricción implica que la totalidad de videos negativos debe ser explorada de una manera adecuada. En lugar de ello, al emplear la definición de *alineamiento de kernels* de la ecuación 2.4 el desbalance inherente a la recolección de las muestras es compensado de manera automática en el algoritmo por medio del uso de la propiedad de centrado de matrices.

3. Se introdujo la condición $\text{Err}_d < \text{Err}_o$ antes de realizar la división de un segmento temporal: de acuerdo con el algoritmo propuesto en (Ryoo and Matthies, 2013) la decisión de crear una división en un segmento temporal depende de que el kernel asociado con la nueva estructura alcance un mayor

Data: conjunto etiquetado de videos $\langle x_i, y_i \rangle_{i=1}^{m} \mid x_i \in \mathbb{R}^{T_i \times d}$, conjunto finito de posiciones temporales $\mathcal{T} : \{t \in \mathcal{T} \mid 0 \leqslant t \leqslant 1\}$.
Result: estructura de la actividad S
Inicializar estructura de la actividad;
while *Existan nodos candidatos* S **do**
  Listar nodos candidatos en S;
  for *nodos candidatos* **do**
    Enumerar posibles estructuras del nodo candidato;
    Calcular $A_L(K_S)$ de todas las estructuras;
  end
  $S' \leftarrow argmax_S A_L(K_S)$;
  nodo explorado $\leftarrow max_S A_L(K_S)$;
  if $Err_d < Err_o$ **then**
    $S \leftarrow S'$ Marcar nodo explorado en S como nodo no terminal;
  else
    Marcar nodo explorado en S como nodo terminal;
  end
  Excluir nodo explorado de la lista de nodos candidatos;
end

Algorithm 1: Pseudocódigo del algoritmo de aprendizaje de la estructura de las actividades.

alineamiento con la función de kernel óptimo, en relación con el alineamiento de la estructura sin realizar particiones. En términos numéricos esta condición es muy sencilla de ser alcanzada, sin garantizar que el alineamiento sea lo suficientemente discriminativo. Por tal motivo, antes de realizar la división de un segmento temporal se verifica que este aumento en la complejidad del modelo garantiza una mejora en términos de reconocimiento. La introducción de esta condición corresponde con el principio de *descripción de mínima longitud*, MDL[8], el cual considera que la mejor hipótesis para un conjunto de observaciones es aquella que permita la mejor compresión de la información.

4. Se propuso una versión competitiva del algoritmo estándar con el fin de que las particiones que se introduzcan contribuyan a la disminución de errores de entrenamiento.

5. Explicación del algoritmo en términos de la similitud basada en kernel y funciones de similitud: a diferencia de la exposición presentada en (Ryoo and Matthies, 2013), la cual involucra de manera indistinta los conceptos de kernel y distancia. En la presentación del algoritmo de descomposición de actividades, únicamente se hace alusión a medidas de similitud, con el fin de evitar confusiones o errores de forma.

2.3. Clasificación de actividades humanas

En la sección anterior se definió en que consiste la estructura de una actividad y como puede ser aprendida a partir de un conjunto etiquetado de videos. Una vez se obtiene la estructura S_a asociada con la actividad a basta con calcular los descriptores del videos de acuerdo con S_a e introducir la ecuación 2.2 dentro del algoritmo de clasificación predilecto. Por simplicidad y debido a que su formulación en términos de la matriz de kernel se ha estudiado de manera amplia (Burges, 1998),(Chang and Lin, 2011), se hizo uso del algoritmo de clasificación basado en máquinas de soporte vectorial para el entrenamiento de los clasificadores de actividades.

[8]El principio de descripción de mínima longitud es una formalización del principio metodológico y filosófico conocido como la *navaja de Ockham*

2.4. Resultados Experimentales

Para validar el algoritmo de aprendizaje discriminativo de la estructura temporal se realizó un experimento con un conjunto de secuencias sintéticos y dos conjuntos de vídeos públicos (Gupta and Davis, 2007),(Messing et al., 2009). Se analizó su desempeño para la tarea de reconocimiento de acciones humanas de forma supervisada y sin introducir ningún tipo de anotación adicional acerca de las sub-acciones que componen una acción.

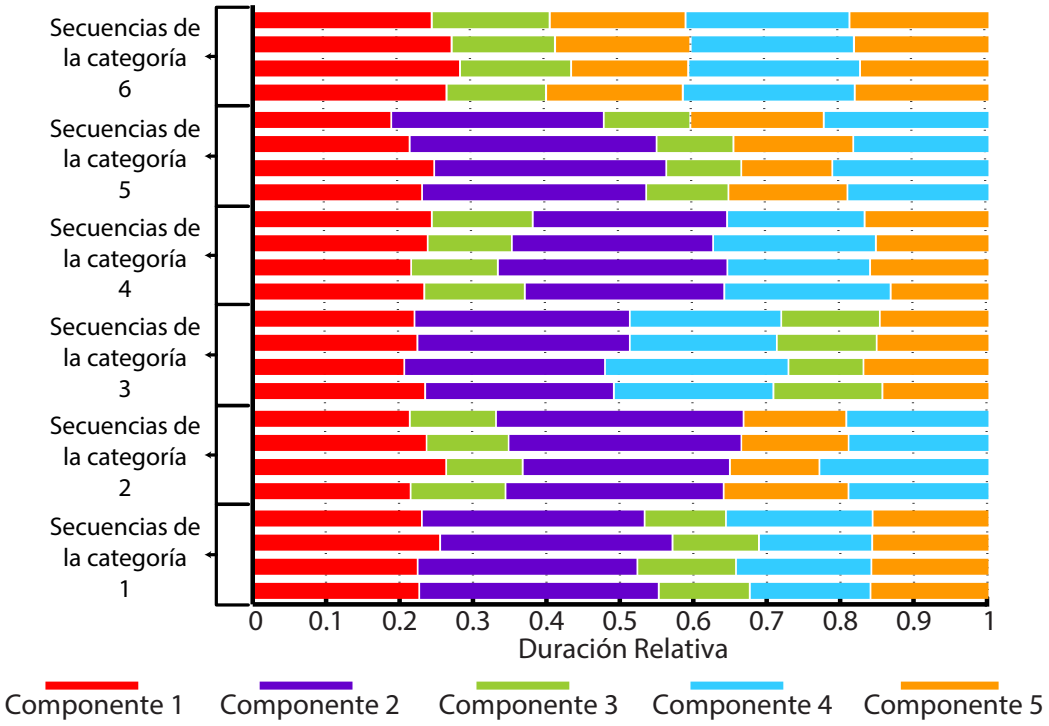


Figura 2.5: Ejemplo de secuencias sintéticas con siete categorías de interés. Los colores representan el tipo de componente observado en un secuencia. Cada color representa la aparición de una componente durante un periodo de tiempo específico. En el texto se encuentran más detalles acerca de la generación de las secuencias

2.4.1. Conjunto de secuencias sintéticas

En primera instancia, se analizó el desempeño del algoritmo de aprendizaje de la estructura temporal de las actividades sobre un conjunto de secuencias sintéticas similares a las que se aprecian en la figura 2.5. De esta forma, se puede constatar

Tabla 2.1: Parámetros de la distribución gaussiana truncada asociada con la duración de cada componente. La duración hace referencia al número de frames en los cuales se aprecia la componente

Componente	μ	σ	Mínimo	Máximo
1	92.0	9.2	27.6	156.4
2	122.0	12.2	36.6	207.4
3	50.0	5.0	15.0	85.0
4	80.0	8.0	24.0	136.0
5	64.0	6.4	19.2	108.8

que el algoritmo es capaz de aprender una representación temporal estructurada, que permite la categorización de secuencias temporales con alta similitud.

Para la generación de las secuencias temporales que se aprecian en la figura 2.5, se definieron las siguientes reglas de producción:

- Categoría 1: Componente 1, Componente 2, Componente 3, Componente 4, Componente 5.
- Categoría 2: Componente 1, Componente 3, Componente 2, Componente 5, Componente 4.
- Categoría 3: Componente 1, Componente 2, Componente 4, Componente 3, Componente 5.
- Categoría 4: Componente 1, Componente 3, Componente 2, Componente 4, Componente 5.
- Categoría 5: Componente 1, Componente 2, Componente 3, Componente 5, Componente 4.
- Categoría 6: Componente 1, Componente 3, Componente 5, Componente 4, Categoría 5.

Con el fin de que las categorías secuenciales sintéticas sean difíciles de distinguir, a la duración de cada categoría se le asoció una distribución de probabilidad gaussiana truncada con los parámetros que se aprecian en la tabla 2.1.

A partir del conjunto de reglas y parámetros mencionados, la generación de una secuencia asignada con la categoría a consiste en el muestreo estadístico de la duración de cada una de las acciones que compone la actividad. La representación de las observaciones en cada frame se realiza a través de un vector de activaciones booleano de dimensión d. Una activación en la i-ésima dimensión del vector indica que la componente d_i ocurre en el frame que se esta observando. d es equivalente al

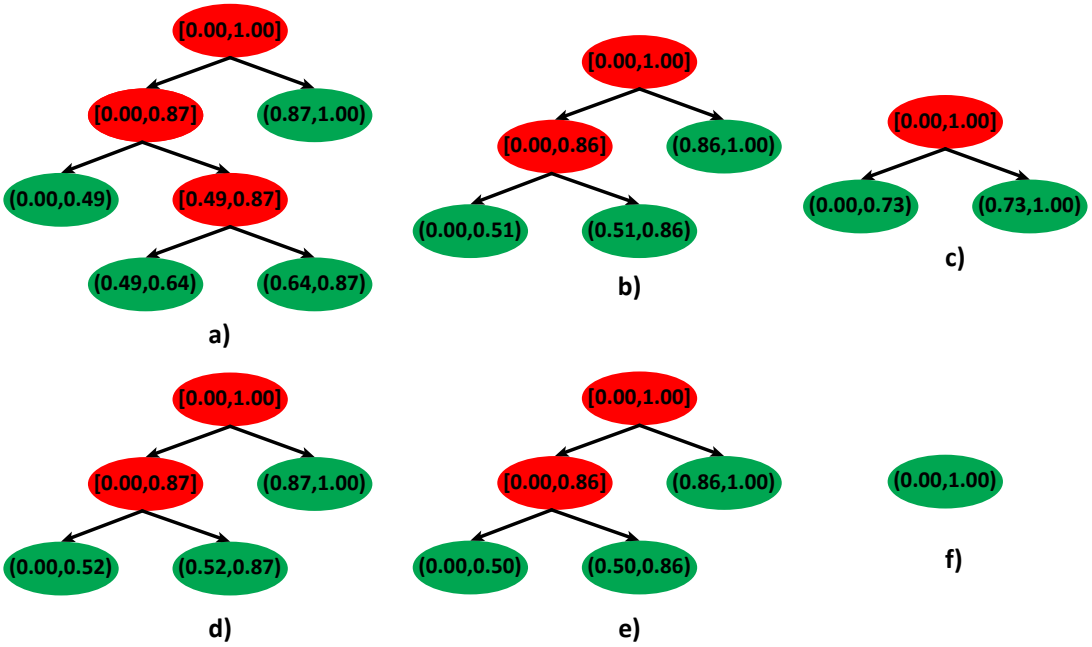


Figura 2.6: Estructuras temporales aprendidas por nuestro algoritmo para el conjunto de secuencias sintéticas de la figura 2.5. Cada categoría de secuencia del 1-5 tiene asociada una estructura temporal (a-e) de manera respectiva. Los nodos verdes representan los segmentos terminales a través de los cuales se representa el contexto temporal de la acción. Mientras que, los nodos rojos representan nodos no-terminales. Al interior de cada nodo se aprecia el intervalo de tiempo relativo asociado con el mismo.

número total componentes a través de las cuales se pueden describir las categorías. Para el ejemplo de la figura 2.5, el número de componentes es igual a 5. La información temporal asociada con una secuencia x_i en un interavo de tiempo (t_1, t_2) se representa mediante un modelo de bolsa de palabras o de activación promedio, en el que cada palabra representa la occurencia de una componente. A partir de la ecuación 2.10 se puede calcular la representación temporal del video x_i en el intervalo de tiempo (t_1, t_2).

$$x_i^{(t_1,t_2)} = \frac{1}{(t_2 - t_1 + 1)} \sum_{t=t_1}^{t_2} x_i^t \tag{2.10}$$

En total se generaron 1200 secuencias, 200 por cada categoría, de las cuales 840 son utilizadas como observaciones de entrenamiento por el algoritmo de aprendizaje de la estructura temporal 2.2 y de clasificación de las categorías. Para validar el desempeño del algoritmo de clasificación de categorías basado en la estructura temporal se utilizaron las 360 secuencias no utilizadas durante entrena-

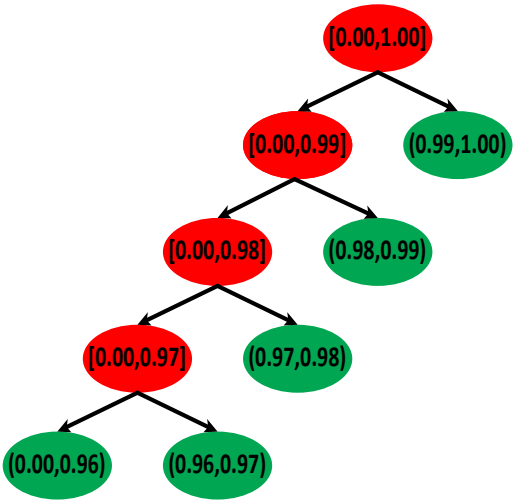


Figura 2.7: Estructura temporal aprendida por el algoritmo Ryoo y Matthies para las secuencias sintéticas etiquetadas con la categoría 5. En comparación con la estructura temporal de la categoría 5 mostrada en la figura 2.6(e), este algoritmo obtiene una estructura densa y con segmentos terminales enfocadas en la descripción temporal fina en lugar de las características dinámicas discriminativas. Las estructuras temporales asociadas con las secuencias sintéticas de las otras categorías son mucho más densas y se presentan en el apéndice A.

miento.

En las figuras 2.6 y 2.7 se aprecian las estructuras temporales discriminativas aprendidas para cada categoría. Los resultados cuantitativos asociados con la clasificación de las secuencias categóricas en términos de una matriz de confusión se aprecian el la figura 2.8.

Nótese que, a pesar de que el número de componentes de las categorías 1-5 es cinco, el máximo número de segmentos temporales terminales identificados por el algoritmo de aprendizaje estructural propuesto es cuatro. La razón de esta discrepancia subyace en que el algoritmo es discriminativo, es decir, durante cada ronda de aprendizaje el algoritmo intenta dividir el intervalo de duración de tal forma que la similitud entre las secuencias asociadas con la categoría de interés y las secuencias irrelevantes sea mínima. A continuación se revisará de manera intuitiva la selección de los tiempos relativos discriminativos al construir la estructura de la *categoría 2*. Por simplicidad se reducirá el problema a la selección del orden adecuado para las particiones $t_1 = 0{,}86, t_2 = 0{,}51$. En este caso se tiene que la partición t_1 permite que las categorías 2 y 5 sean diferentes de las demás. En tanto que la partición t_2 permite que las categorías 1,2,4 y 5 sean diferentes del resto. Asumiendo que se debe escoger una partición, el sentido común optaría por la par-

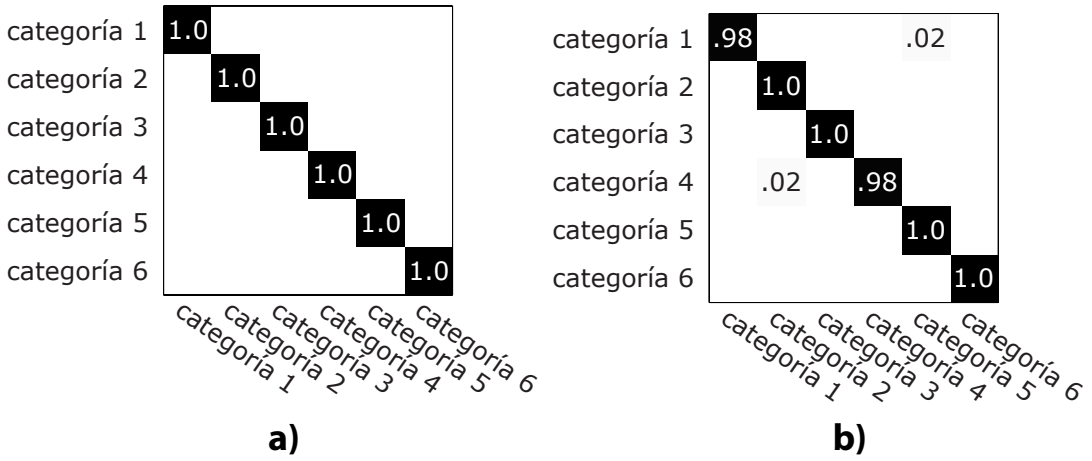


Figura 2.8: Matrices de confusión de las categorías de secuencias sintéticas de la figura 2.5 obtenidas a través de la representación del algoritmo propuesto (a) y por el algoritmo de Ryoo y Matthies (b).

tición t_1 puesto que su capacidad de hacer a la categoría 2 distinguible del resto es mejor. El algoritmo empleado trabaja de manera similar, por tal motivo se denomina discriminativo. Esto implica que la localización de las segmentos temporales terminales que representan a la actividad se escogen con base en el resultado de la clasificación.

Una ventaja de las modificaciones realizadas al algoritmo es que permiten una representación compacta de las actividades, a diferencia del algoritmo original de (Ryoo and Matthies, 2013). Representar las actividades de manera compacta es ideal en la medida en que evita que el algoritmo introduzca particiones que no contribuyen a mejorar el desempeño de clasificación. En otras palabras, se puede considerar que el algoritmo modificado es mesurado en relación con el original. La importancia de utilizar un algoritmo mesurado se puede apreciar al comparar la representación de la *categoría 6*. Esta categoría es distinguible del resto al hacer uso de un modelo global, puesto que no contiene el componente 2. En ese caso, el algoritmo original de Ryoo y Matthies complica la representación de la actividad debido a que únicamente hace uso de la medida de alineamiento para introducir divisiones, desembocando en una representación de la misma categoría en término de 17 segmentos terminales.

En términos cuantitativos, el desempeño del algoritmo con todas las modificaciones realizadas en la tarea de clasificación es del 100 % demostrando las bondades de una adecuada representación de la estructura temporal de las actividades. Nótese que, el desempeño del algoritmo original de (Ryoo and Matthies, 2013)

Tabla 2.2: Desempeño de clasificación de diferentes modelos basados en segmentos rígidos que capturan la evolución temporal de las actividades

Estructuras de reconocimiento dínamicas	Precisión (%)
Nuestro Algoritmo estructural	100.00
Algoritmo estructural de (Ryoo and Matthies, 2013)	99.44
1-Segmento temporal	34.17
2-Segmentos temporales uniformes	53.89
3-Segmentos temporales uniformes	92.78
4-Segmentos temporales uniformes	100.00
5-Segmentos temporales uniformes	100.00
6-Segmentos temporales uniformes	100.00

es ligeramente inferior. Una posible causa de ello, sea que el excesivo número de segmentos temporales cause un *sobreajuste* provocando que la capacidad de generalización se vea comprometida. Asimismo, en la tabla 2.2 se resumen el desempeño de clasificación de nuestro algoritmo con respecto a otro modelos de estructura fija que capturan la evolución temporal de las actividades. A diferencia de los modelos basados en más de cuatro segmentos uniformes, las estructuras temporales determinas por los *algoritmos de descomposición de actividades* son particulares para cada categoría y pueden ser más compactas aplicando las modificaciones subrayadas en la sección 2.2.

Para culminar el análisis del desempeño del algoritmo sobre las secuencias sintéticas. Se estudió el efecto de las modificaciones realizadas al algoritmo original propuesto en (Ryoo and Matthies, 2013). Para ello, se utilizó un conjunto de siete actividades con las características que resume la tabla 2.3. Nótese que, en este experimento la distribución de probabilidad de cada componente es idéntica con el fin de estudiar los efectos de una mayor variabilidad en la duración de las mismas.

En la figura 2.9 se resumen los resultados del experimento. Para garantizar la validez estadística de los resultados, cada corrida se repitió cincuenta veces. En la figura 2.9 se reporta el desempeño promedio de cada algoritmo agrupados de acuerdo con el nivel de variabilidad de cada componente.

Los resultados obtenidos confirman las ventajas de las modificaciones realizadas al algoritmo de Ryoo y Matthies (Ryoo and Matthies, 2013). Por ejemplo, es notable como el algoritmo propuesto mantiene el desempeño de clasificación de las categorías, aún cuando la razón media sobre desviación estándar es del 30 %. Asimismo, llama la atención que cuando la similitud en el orden de las categorías es muy alta el algoritmo en su versión competitiva en lugar de favorecer el *sobreajuste*

Tabla 2.3: Características de las secuencias sintéticas utilizadas para contrastar las modificaciones realizadas al algoritmo de Ryoo y Matthies.

Categoría	Componentes
1	1, 2, 3, 4
2	1, 2, 4, 3
3	2, 1, 3, 4
4	1, 3, 2, 4
5	3, 1, 2, 4
6	2, 1, 4, 3
7	3, 1, 4, 2

Componente	μ	σ	Mínimo	Máximo
1	40.0	0.0, 2.0, 4.0, 6.0, 8.0, 10.0, 12.0	12.0	68.0
2	40.0	0.0, 2.0, 4.0, 6.0, 8.0, 10.0, 12.0	12.0	68.0
3	40.0	0.0, 2.0, 4.0, 6.0, 8.0, 10.0, 12.0	12.0	68.0
4	40.0	0.0, 2.0, 4.0, 6.0, 8.0, 10.0, 12.0	12.0	68.0

Número total de secuencias	1400
Número de secuencias por actividad	200
Porcentaje de secuencias utilizadas en entrenamiento	70 %

permite que se obtengan modelos estructurales efectivos.

En relación con las modificaciones relacionadas con la función de kernel óptimo y la matriz de kernel, se puede asegurar que la propiedad de centrado contribuye en gran medida al desempeño exitoso del algoritmo. Por otro lado, se evidencia que el uso de nuestra función de kernel es destacable en 3 de los 8 casos de estudio, y se desempeña a la par de la función de Ryoo y Matthies en en el resto de los casos.

Por último, vale la pena considerar el resultado obtenido cuando se hace uso de la condición basada en el principio de descripción de mínima longitud. En general se podría decir que su inclusión afecta ligeramente el desempeño del algoritmo de original. Sin embargo, es importante recordar que su contribución en el algoritmo es mantener una representación simple y evitar estructuras temporales excesivamente complejas. En ese sentido, ese pequeño porcentaje que se pierde no parece tan relevante, cuando se obtienen estructuras temporales compactas como las presentadas de la figura 2.6.

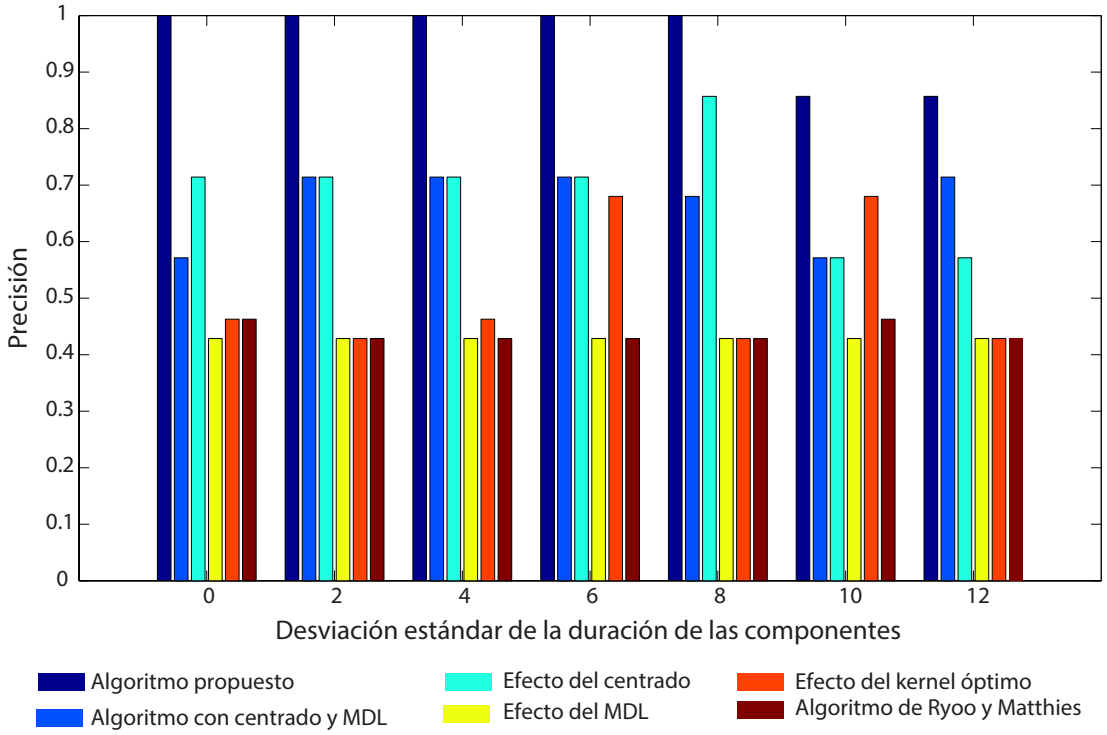


Figura 2.9: Efecto de las modificaciones realizadas al algoritmo de Ryoo y Matthies sobre el segundo conjunto de secuencias sintéticas. (Se interpreta mejor a color)

2.4.2. Conjunto de acciones de Gupta

Con el fin de analizar el desempeño del algoritmo de descomposición de actividades en videos reales y demostrar extender su uso al dominio de la descripción de las interacciones entre humanos y objetos, se revisó su desempeño sobre el conjunto de videos públicos de Gupta *et. al* (Gupta and Davis, 2007) para la tarea de reconocimiento de acciones.

Para la evaluación del algoritmo, se utilizo el método de validación *dejando uno por fuera*, LOO[9], con el fin de estudiar la capacidad de generalización del algoritmo. En la práctica, únicamente fue posible utilizar cuatro sujetos de prueba bajo esta metodología. Debido a que los demás actores no contaban con videos de todas las acciones propuestas por los autores.

Para el aprendizaje de la estructura temporal asociada con cada acción se empleó el descriptor de interacción de posición relativa entre la persona y el objeto ϕ_l, descrito en la sección 1.2. De esta forma, al aprender la estructura temporal

[9]Acrónimo tomado del término en inglés “Leave One Out”

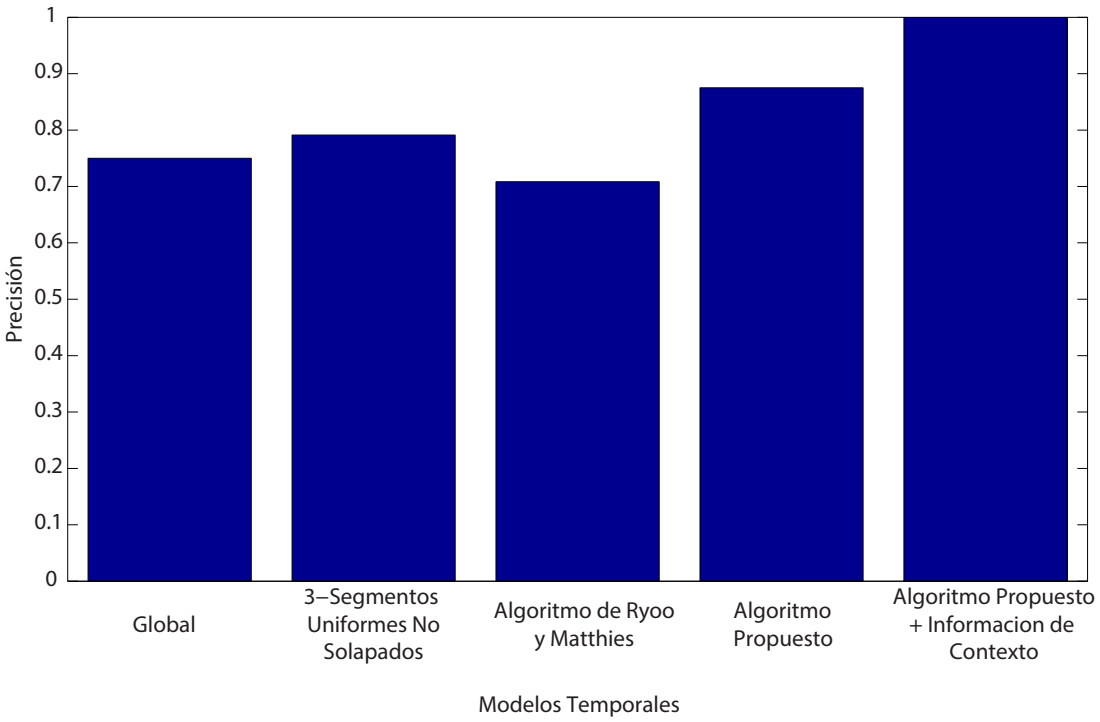


Figura 2.10: Desempeño de las estructuras temporales usando descriptores que representan la interacción espacial entre humanos y objetos en el conjunto de videos de Gupta.

asociada con cada acción se determina un estructura temporal discriminativa que captura como se llevan a cabo las interacciones espaciales entre la persona y el objeto. Para el entrenamiento del algoritmo que permite la clasificación de las acciones humanas se empleó el procedimiento descrito en la sección 1.4.

De acuerdo con la versión original del algoritmo (Ryoo and Matthies, 2013) se empleó una función de kernel χ^2. Debido al pequeño número de datos de entrenamiento, fue necesario computar el descriptor de interacciones espaciales ϕ_l utilizando las versiones originales de los datos de entrenamiento y versiones reflejadas de manera horizontal. En la figura 2.10 se aprecia el desempeño de los diferentes esquemas que describen la evolución temporal de las interacciones entre humanos y objetos. Llama la atención que el algoritmo de Ryoo y Matthies se desempeñe ligeramente por debajo de una descripción que resuma la información de las interacciones de manera global. Sin embargo, al realizar las modificaciones propuestas y encontrar las actividades de manera competitiva se encuentra que adquiere un mejor que la representación uniforme basada en tres segmentos no solapados. De hecho, cuando se complementa la información de las interacciones

espacio-temporales entre humanos y objetos con la información del contexto del objeto asociado con la acción, el resultado de la clasificación es perfecto.

Por otro lado, en la figura 2.11 se compara el efecto de las modificaciones realizadas al algoritmo de (Ryoo and Matthies, 2013). Aquí se aprecia que cada una de las modificaciones propuestas impacta de manera positiva en el desempeño de clasificación. De forma similar a como se obtuvo en el experimento con secuencias sintéticas, las ventajas más grandes se obtienen empleando el centrado de las matrices y un esquema de aprendizaje competitivo.

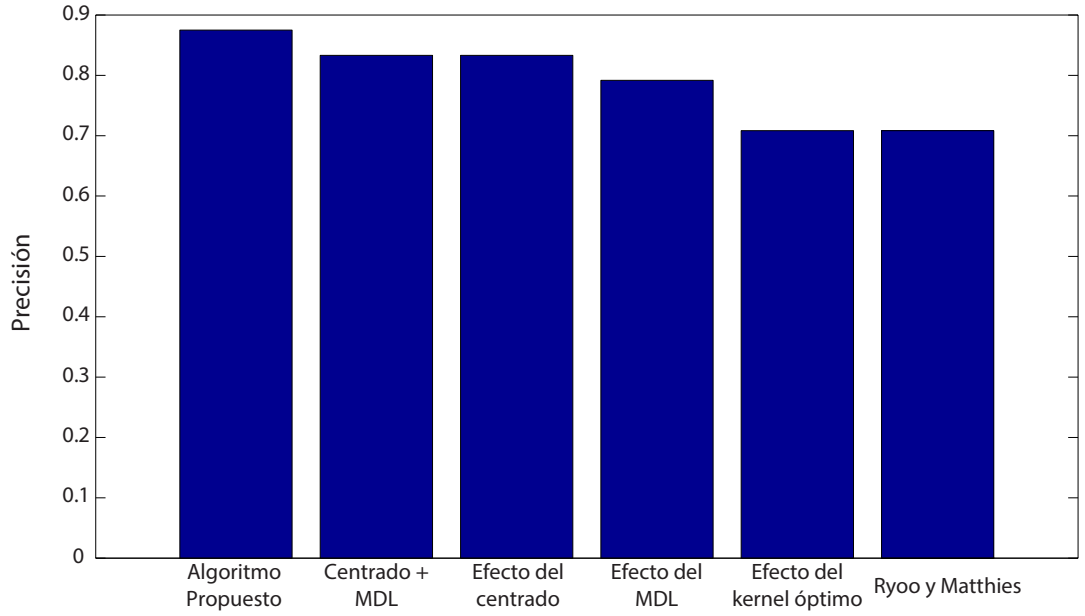


Figura 2.11: Efecto de las modificaciones realizadas al algoritmo de Ryoo y Matthies sobre el conjunto de videos de Gupta.

En las figura 2.12 se aprecian las estructuras temporales de las acciones del conjunto de videos de Gupta aprendidas por la modificación propuesta para el algoritmo de Ryoo y Matthies y por este algoritmo también. De manera similar a lo apreciado con las secuencias sintéticas, el algoritmo de (Ryoo and Matthies, 2013) produce estructuras más densas. Esto se debe a que la condición empleada para dividir un segmento no es lo suficientemente robusta. Por el contrario, con la condición basada en el principio de *descripción de mínima longitud*, MDL, se obtienen estructuras temporales más compactas y efectivas. A modo de comparación en la tabla 2.4 se resumen las características de las estructuras temporales obtenidas.

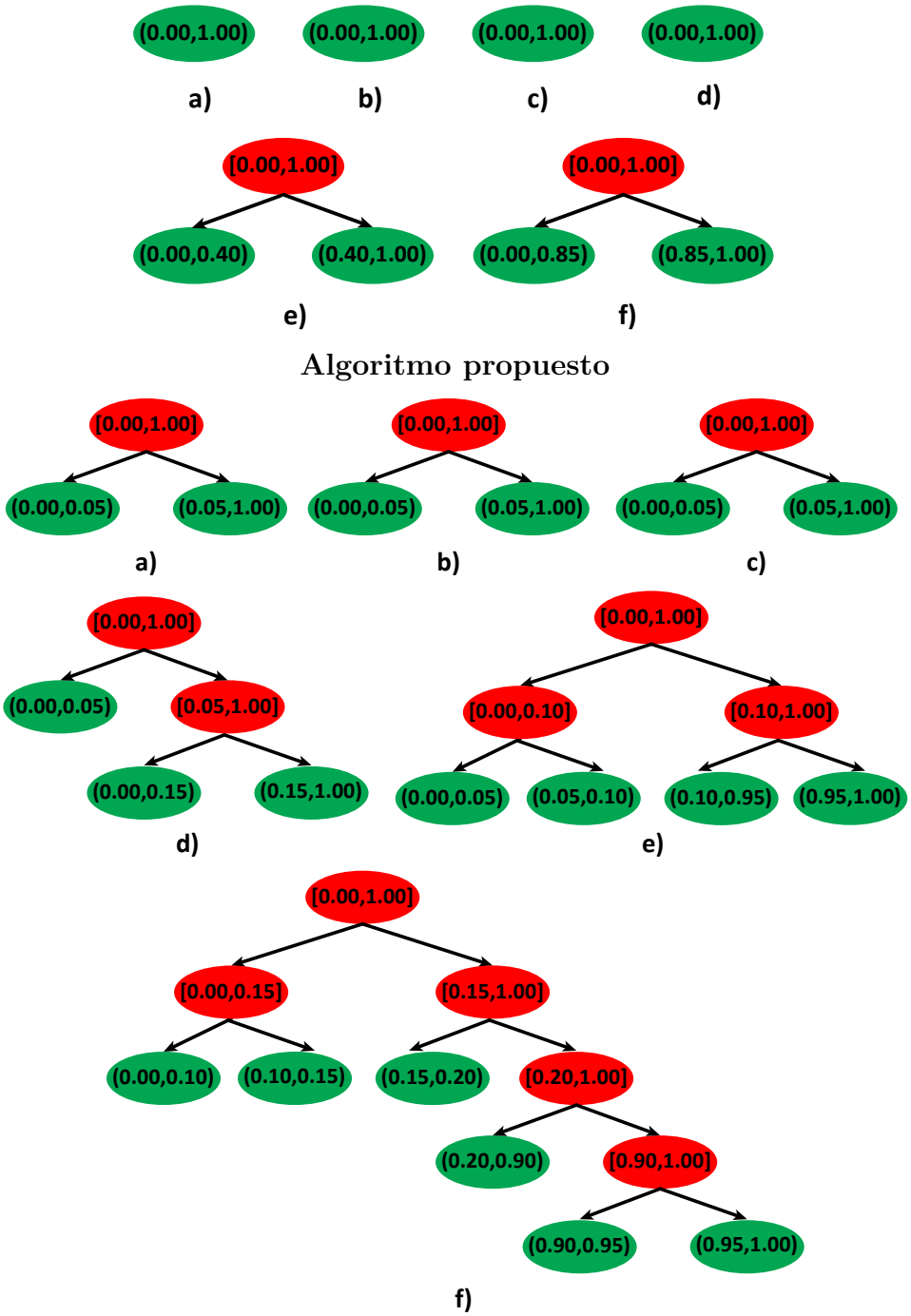


Figura 2.12: Estructuras temporales de las acciones del conjunto de videos de Gupta aprendidas, de (a-f) se muestran las estructuras de las acciones llamar por teléfono, contestar el teléfono, beber, encender una linterna, servir de una taza y rociar spray, respectivamente.

Tabla 2.4: Comparación cuantitativa de las estructuras temporales aprendidas aprendidas a partir del algoritmo de Ryoo y Matthies y con las modificaciones realizadas

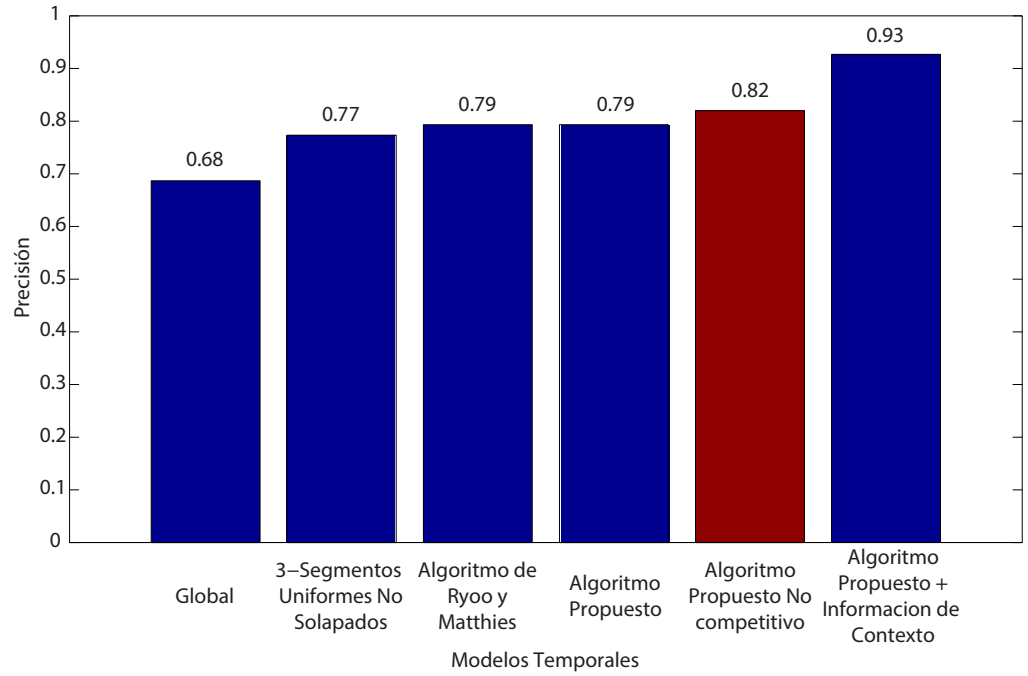
Llamar por teléfono	No. segmentos terminales	No. nodos	Profundidad
Algoritmo de Ryoo y Matthies	2	3	1
Algoritmo Propuesto	1	1	0
Contestar el teléfono	**No. segmentos terminales**	**No. nodos**	**Profundidad**
Algoritmo de Ryoo y Matthies	2	3	1
Algoritmo Propuesto	1	1	0
Beber	**No. segmentos terminales**	**No. nodos**	**Profundidad**
Algoritmo de Ryoo y Matthies	2	3	1
Algoritmo Propuesto	1	1	0
Encender una linterna	**No. segmentos terminales**	**No. nodos**	**Profundidad**
Algoritmo de Ryoo y Matthies	3	5	2
Algoritmo Propuesto	1	1	0
Servir de una taza	**No. segmentos terminales**	**No. nodos**	**Profundidad**
Algoritmo de Ryoo y Matthies	7	4	2
Algoritmo Propuesto	2	3	1
Rociar spray	**No. segmentos terminales**	**No. nodos**	**Profundidad**
Algoritmo de Ryoo y Matthies	6	11	4
Algoritmo Propuesto	2	3	1

2.4.3. Conjunto de acciones de la vida cotidiana de Rochester

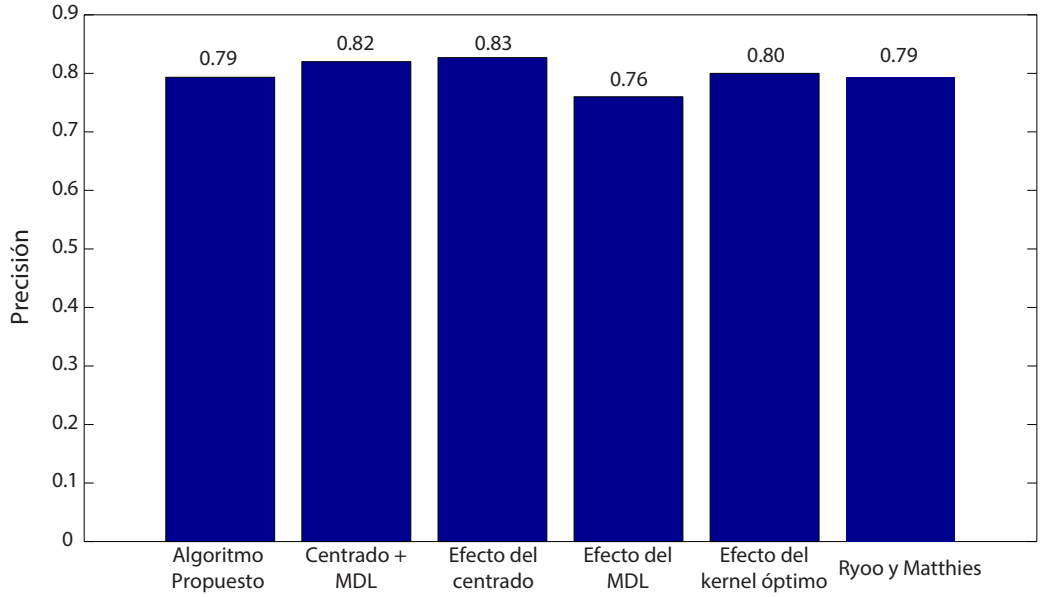
Para complementar los resultados obtenidos en los otros dos experimentos, se clasificaron las acciones humanas presentes en el conjunto de videos de (Messing et al., 2009) empleando el descriptor espacial de las interacciones entre humanos y objetos presentado en la sección 1.2, bajo el mismo esquema metodológico discutido en la subsección anterior.

De forma similar a los resultados obtenidos anteriormente, en este conjunto de videos también se logra apreciar la ventaja de una representación temporal adecuada de las acciones humanas. En la figura 2.13a se aprecia el desempeño del algoritmo propuesto en relación con otras estructuras que calculan la evolución temporal de las interacciones entre humanos y objetos.

Por otro lado, al comparar el efecto de las modificaciones realizadas al algoritmo de (Ryoo and Matthies, 2013) se obtuvieron los resultados que se aprecian en la figura 2.13b. En este caso, se aprecia que el desemepño del algoritmo pro-



(a)



(b)

Figura 2.13: Desempeño de clasificación de los modelos temporales a traves de la descripcion dinamica de las interacciones entre humanos y objetos (a). Efecto de las modificaciones realizadas al algoritmo de Ryoo y Matthies sobre el conjunto de videos de Rochester.

puesto es equivalente al obtenido por Ryoo y Matthies. La explicación de esto se aprecia en la misma figura, al observar el desempeño del algoritmo propuesto sin un esquema de entrenamiento competitivo. Al parecer, el carácter competitivo del algoritmo esta causando un *sobreajuste* que provoca que el desempeño disminuya ligeramente, sin ser inferior al del algoritmo de (Ryoo and Matthies, 2013). Vale la pena mencionar nuevamente, la importancia de centrar las matrices de kernel antes de realizar el alineamiento para obtener un buen desempeño. En futuros trabajos, se explorarán esquemas de entrenamiento más elaborados basados en la teoría de boosting que permitan obtener un comportamiento competitivo sin sufrir de *sobreajuste*.

Conclusión

En este trabajo se presentó una metodología algorítmica que permite el reconocimiento de las acciones humanas en videos a partir de las interacciones entre humanos y objetos. Esto se logró a partir de la representación de las acciones humanas en términos de las relaciones espacio temporales entre un humano y un objeto (capítulo 1, 2). En relación con los trabajos anteriores, este trabajo exploró y se demostró las bondades de las representaciones estructurales que tienen en cuenta el contexto temporal y dinámico de las interacciones entre humanos y objetos a lo largo del video (Escorcia and Niebles, 2013).

En primer lugar, se mostró cómo a partir de una adecuada representación de las interacciones dinámicas entre humanos y objetos se puede capturar la evolución temporal de las relaciones entre los mismos (capítulo 1). A partir de la descripción semántica propuesta se apreciaron resultados competitivos en relación con los descriptores del estado del arte en interacciones entre humanos y objetos en dos conjuntos de videos públicos (sección 1.2, 1.3, 1.4). Por otro lado, se demostró la ventaja en términos del desempeño de clasificación que producen el modelado de las características dinámicas de las interacciones entre humanos y objetos. De esta manera, los resultados obtenidos por nuestra representación dinámica de las interacciones sobrepasaron los resultados que hasta ese momento ostentaban el estado del arte en los dos conjuntos de videos públicos evaluados (Escorcia and Niebles, 2013).

A futuro, es deseable explorar la manera de integrar la descripción de las interacciones espacio temporales con el contexto estructurado entre múltiples objetos y acciones. Asimismo, es importante permitir que la descripción de las interacciones permita enfocarse en el objeto como actor principal de la escena. Ésto permitirá reconocer acciones y actividades como cocinar, en las cuales el interés de la acción se concentra en la manipulación y el aspecto de los utensilios y los alimentos.

En segunda instancia se enunciaron las desventajas de las representaciones temporales estrictas, las cuales definen de manera arbitraria el número de subacciones que componen una actividad (sección 2.1). Con base en éstas, se analizó y

se empleó el algoritmo de (Ryoo and Matthies, 2013) con el fin de describir de manera flexible la estructura espacio-temporal de las interacciones entre humanos y objetos (capítulo 2). A partir de los resultados cuantitativos y cualitativos del algoritmo de Ryoo y Matthies, se propusieron algunas modificaciones al mismo con el fin de evitar una sobre-representación de las estructuras temporales asociadas con las actividades (sección 2.2). Las modificaciones realizadas permiten obtener un mejor desempeño a nivel cuantitativo y cualitativo para la clasificación de secuencias sintéticas y la clasificacion de acciones humanas la mayoría de los casos (sección 2.4).

En futuros trabajos se planea explorar si al introducir flexibilidad en la ubicación y duración de cada uno de los segmentos temporales, se mejora el reconocimiento de las actividades. Estos resultados, se complementarán con un estudio del algoritmo de descomposición de actividades propuesto en videos con pocas restricciones. Los cuales son representados a partir de los descriptores representativos, basados en características fotométricas de bajo nivel.

Apéndice A

Estructuras temporales sintéticas

A continuación se visualizan las estructuras temporales aprendidas mediante el algoritmo de Ryoo y Matthies para el primer experimento en secuencias sintéticas. Las imágenes presentadas a continuación fueron obtenidas a partir del código fuente con el que se obtuvieron los experimentos y debido a su profundidad no fue posible realizarlas a mano en un programa de gráficos vectoriales. Si se desean revisar con mayor detenimiento, es posible hacerlo a través de los siguientes URL:

`https://www.dropbox.com/s/4z1wjj5pyp3ht19/ryoo_actv1.png`

`https://www.dropbox.com/s/4tggvgbo2hc12du/ryoo_actv2.png`

`https://www.dropbox.com/s/jv6htr5qsg9fenk/ryoo_actv3.png`

`https://www.dropbox.com/s/t6m388addtwwra8/ryoo_actv4.png`

`https://www.dropbox.com/s/19t1zvmxx8ugqvn/ryoo_actv6.png`

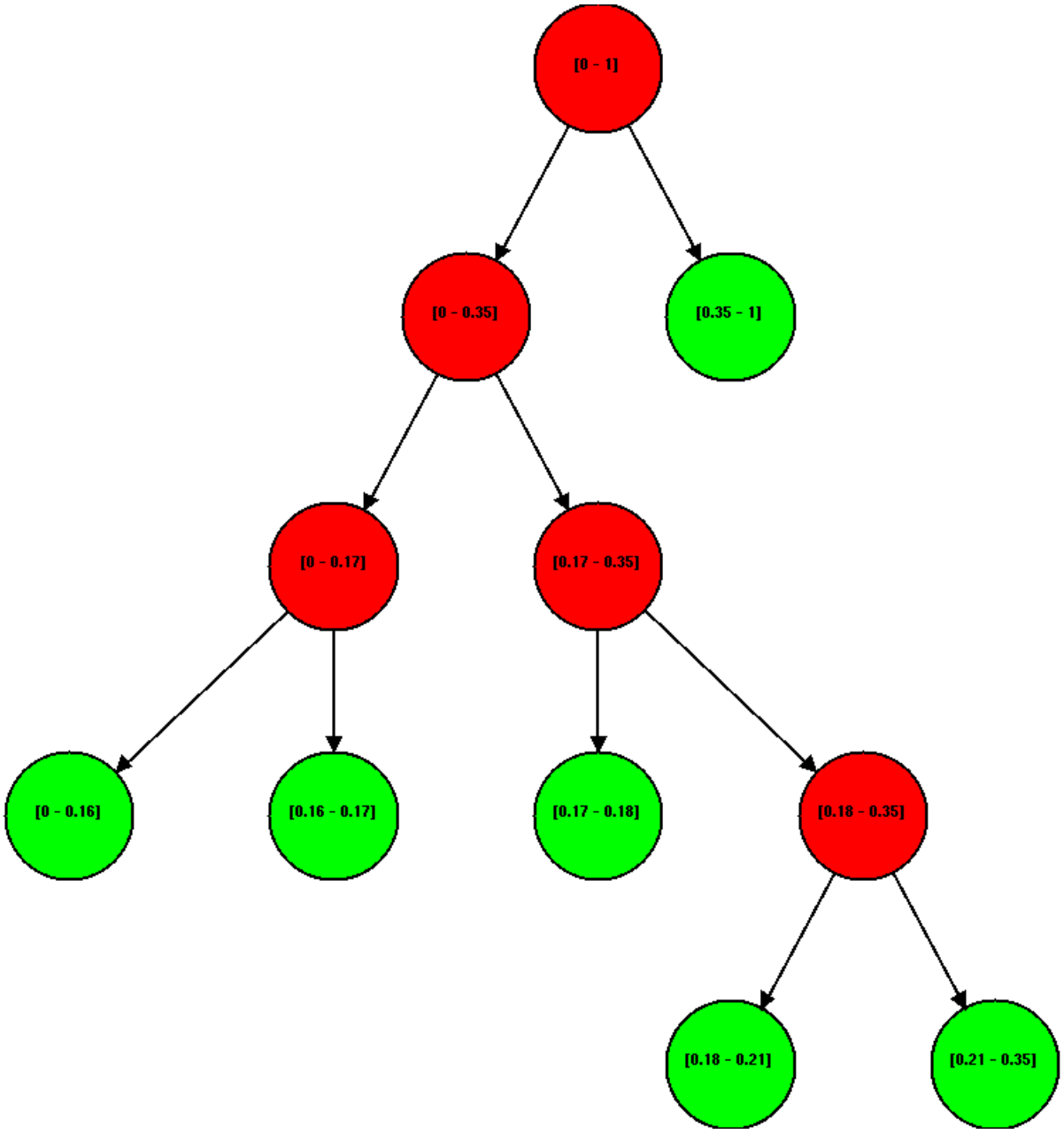


Figura A.1: Estructura para la categoría de secuencia sintética 1 de la figura 2.5

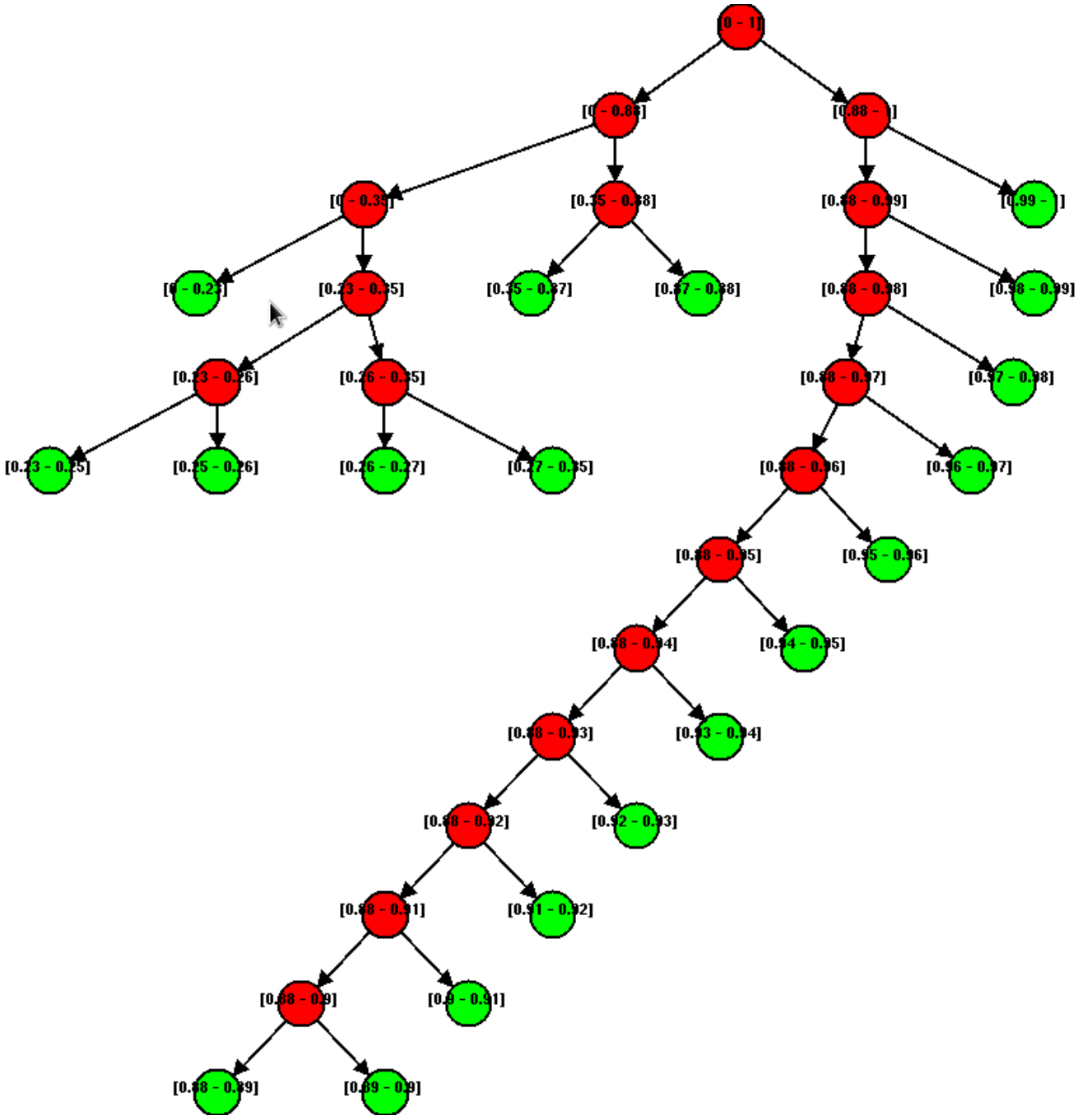


Figura A.2: Estructura para la categoría de secuencia sintética 2 de la figura 2.5

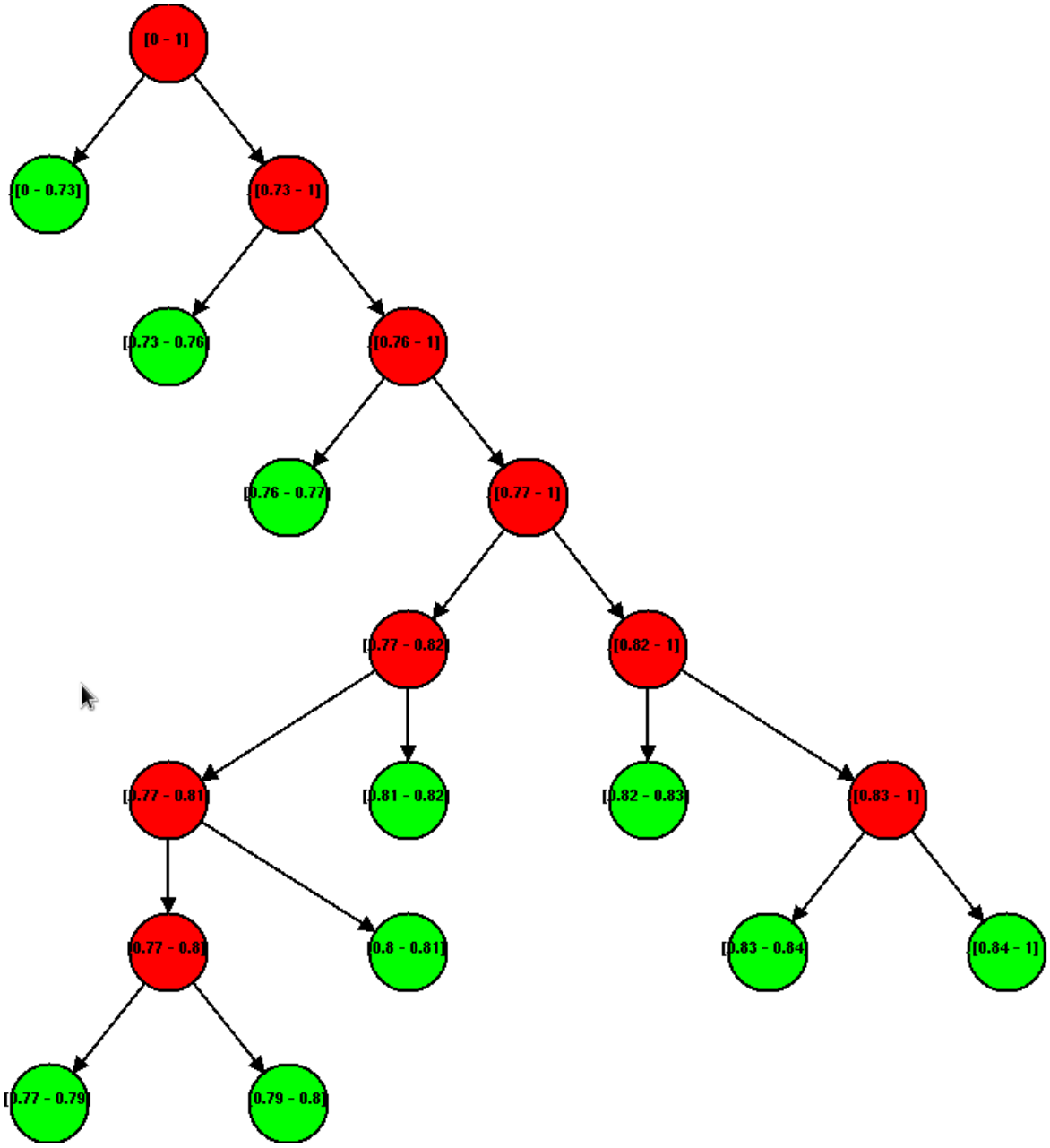


Figura A.3: Estructura para la categoría de secuencia sintética 3 de la figura 2.5

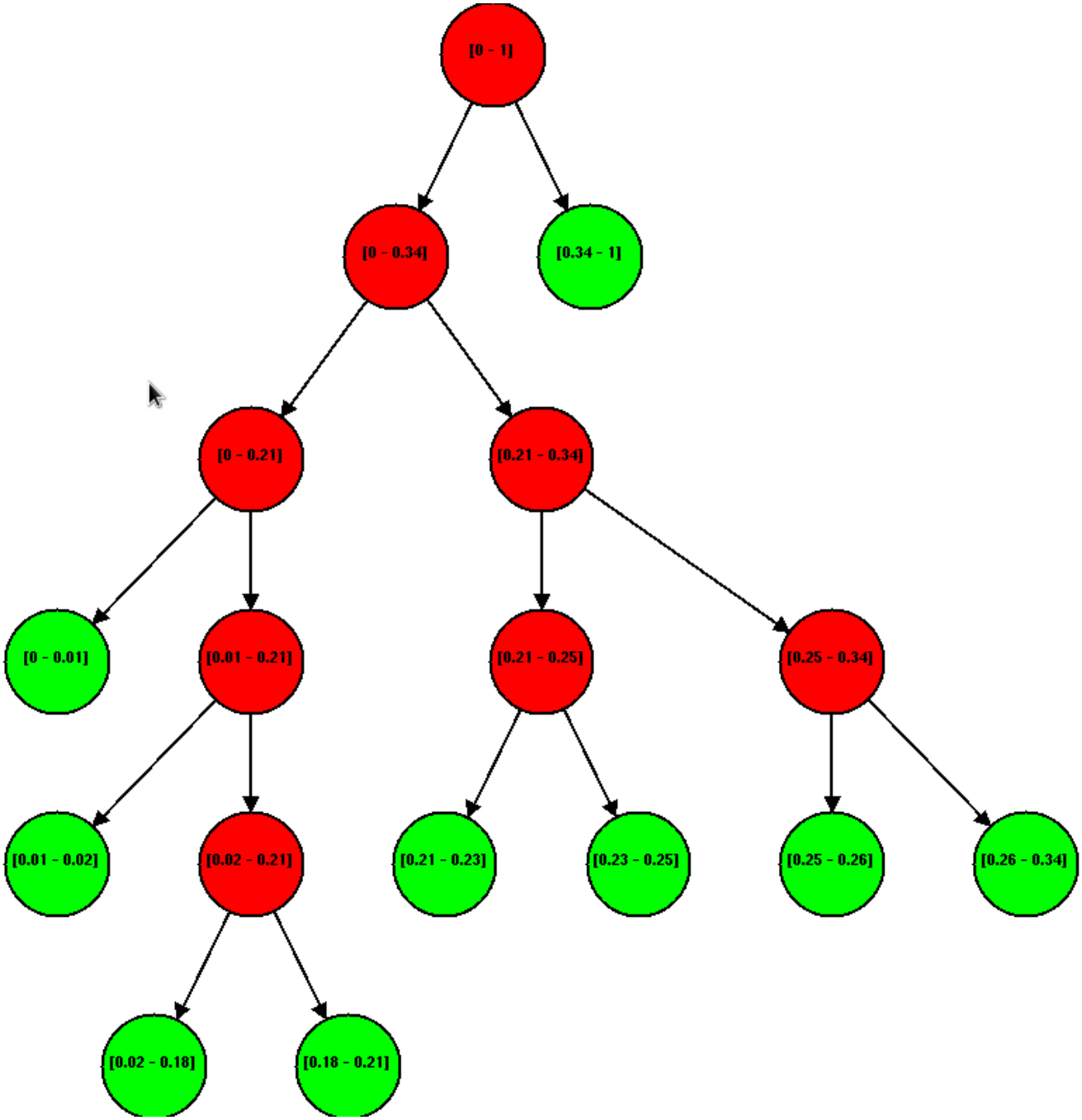


Figura A.4: Estructura para la categoría de secuencia sintética 4 de la figura 2.5

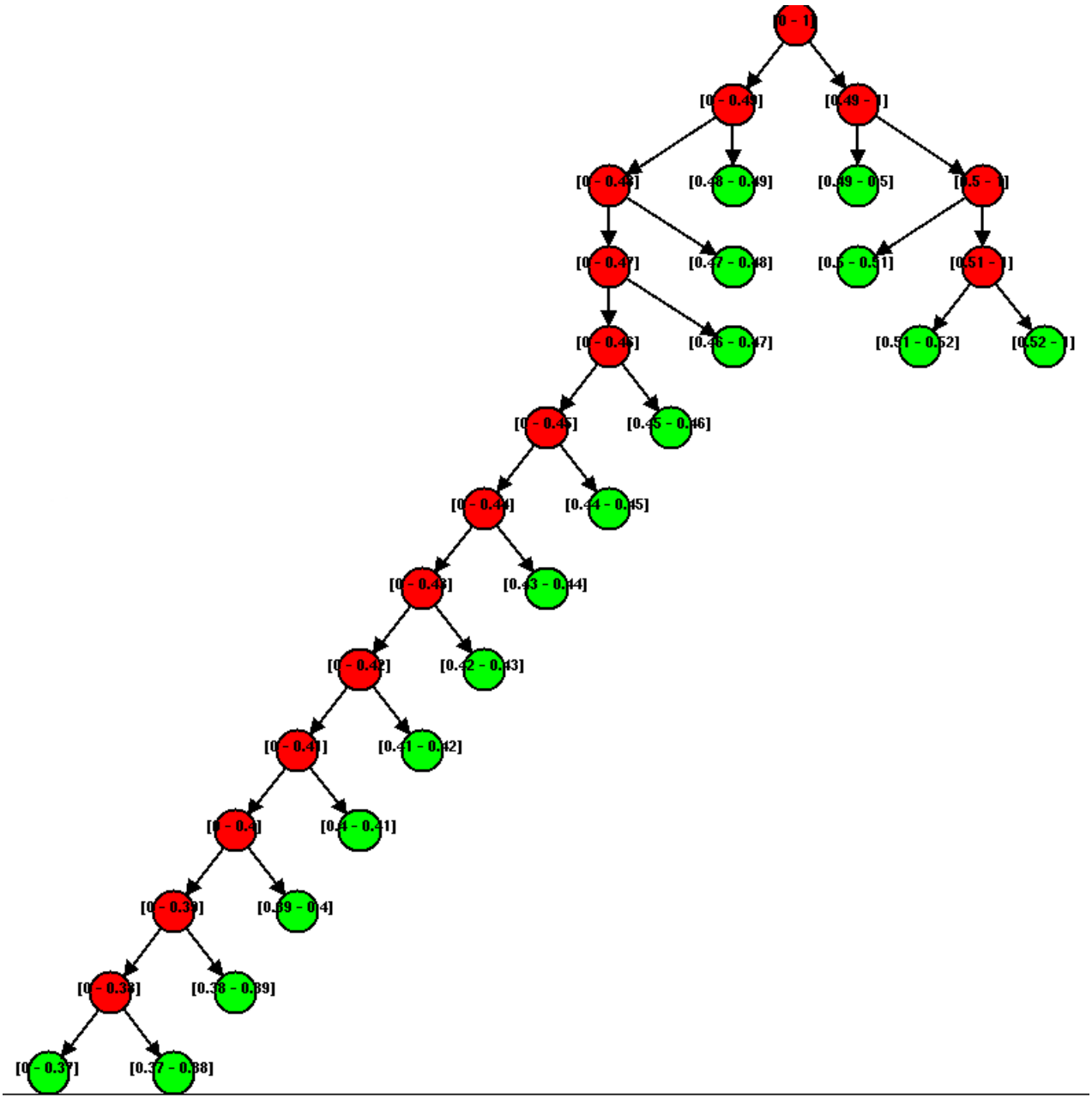


Figura A.5: Estructura para la categoría de secuencia sintética 6 de la figura 2.5

BibliografBibliograf

Bibliograf

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis. *ACM Computing Surveys*, 43(3):1–43.

Barbic, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J. K., and Pollard, N. S. (2004). Segmenting motion capture data into distinct behaviors. In *Graphics Interface*, pages 185–194.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates.

Bourdev, L., Maji, S., Brox, T., and Malik, J. (2010). Detecting People Using Mutually Consistent Poselet Activations. In *Proceedings of European Conference on Computer Vision, ECCV*, Lecture Notes in Computer Science. Springer.

Brendel, W. and Todorovic, S. (2011). Learning spatiotemporal graphs of human activities. In *International Conference on Computer Vision, ICCV*.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. S. (2001). On kernel-target alignment. In *Advances in Neural Information Processing Systems, NIPS*, pages 367–373.

Delaitre, V., Laptev, I., and Sivic, J. (2010). Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference, BMVC*.

Desai, C. and Ramanan, D. (2012). Detecting actions, poses, and objects with relational phraselets. In *Proceedings of European Conference on Computer Vision, ECCV*.

Desai, C., Ramanan, D., and Fowlkes, C. (2009). Discriminative models for multi-class object layout. In *IEEE International Conference on COmputer Vision, ICCV*.

Desai, C., Ramanan, D., and Fowlkes, C. (2010). Discriminative models for static human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPRW*.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior Recognition via Sparse Spatio-Temporal Features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*.

Escorcia, V. and Niebles, J. C. (2013). Spatio-temporal human-object interactions for action recognition in videos. In *IEEE International Conference on Computer Vision, ICCV Workshops*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Gaidon, A., Harchaoui, Z., and Schmid, C. (2011). Actom Sequence Models for Efficient Action Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Gaidon, A., Harchaoui, Z., and Schmid, C. (2012). Recognizing activities with cluster-trees of tracklets. In *Procedings of the British Machine Vision Conference, BMVC*, pages 1–13.

Gupta, A. and Davis, L. S. (2007). Objects in Action: An Approach for Combining Action Understanding and Object Perception. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gupta, A., Kembhavi, A., and Davis, L. (2009). Observing human-object interactions: Using spatial and functional compatibility for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1775–1789.

Ikizler, N. and Forsyth, D. A. (2008). Searching for Complex Human Activities with No Visual Examples. *International Journal of Computer Vision*, 80(3):337–357.

Ivanov, Y. A. and Bobick, A. F. (2000). Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):852–872.

Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., and Shah, M. (2012). High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, pages 1–29.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Prentice

Hall, 2 edition.

Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *International Journal Robotic Research*, 32(8):951–970.

Lan, T., Wang, Y., and Mori, G. (2011). Discriminative figure-centric models for joint action localization and recognition. In *IEEE International Conference on Computer Vision, ICCV*.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Laxton, B., Lim, J., and Kriegman, D. J. (2007). Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Liu, J., Kuipers, B., and Savarese, S. (2011). Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Messing, R., Pal, C., and Kautz, H. (2009). Activity Recognition using the velocity histories of tracked keypoints. In *IEEE International Conference on Computer Vision, ICCV*.

Moore, D. J. and Essa, I. A. (2002). Recognizing multitasked activities from video using stochastic context-free grammar. In *Proceedings of Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence*, pages 770–776.

Nater, F., Grabner, H., and Gool, L. J. V. (2011). Temporal relations in videos for unsupervised activity analysis. In *Procedings of the British Machine Vision Conference, BMVC*, pages 1–11.

Nguyen, M. H., Lan, Z.-Z., and la Torre, F. D. (2011). Joint segmentation and classification of human actions in video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3265–3272.

Niebles, J. C., Chen, C.-w., and Fei-fei, L. (2010). Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In *Proceedings of European Conference on Computer Vision, ECCV*, volume 6312 of *Lecture Notes in Computer Science*, pages 1–14. Springer.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision, IJCV*, 79(3):299–318.

Patron-Perez, A., Marszalek, M., Reid, I., and Zisserman, A. (2012). Structured

learning of human interactions in tv shows. *Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453.

Pei, M., Jia, Y., and Zhu, S.-C. (2011). Parsing video events with goal inference and intent prediction. In *IEEE International Conference on Computer Vision, ICCV*, pages 487–494.

Prest, A., Ferrari, V., and Schmid, C. (2013). Explicit modeling of human-object interactions in realistic videos. *IEEE transactions on pattern analysis and machine intelligence*, pages 9–16.

Prest, A., Leistner, C., Civera, J., Schmid, C., and Ferrari, V. (2012). Learning Object Class Detectors from Weakly Annotated Video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '08, pages 1–8. IEEE.

Ryoo, M. S. and Matthies, L. (2013). First-person activity recognition: What are they doing to me? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2730–2737.

Sadanand, S. and Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Sadeghi, M. A. and Farhadi, A. (2011). Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shah, M. (2009). Recognizing realistic actions from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Si, Z., Pei, M., Yao, B. Z., and Zhu, S.-C. (2011). Unsupervised learning of event and-or grammar and semantics from video. In *IEEE International Conference on Computer Vision, ICCV*, pages 41–48.

Singh, S., Gupta, A., and Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In *Proceedings of European Conference on Computer Vision, ECCV*.

Tang, K., Fei-Fei, L., and Koller, D. (2012a). Learning Latent Temporal Structure for Complex Event Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Providence, RI, USA.

Tang, K., Ramanathan, V., Fei-Fei, L., and Koller, D. (2012b). Shifting Weights: Adapting Object Detectors from Image to Video. In *Advances in Neural Information Processing Systems, NIPS*, pages 1–9.

Tran, S. D. and Davis, L. S. (2008). Event modeling and recognition using markov logic networks. In *Proceedings of European Conference on Computer Vision, ECCV*, pages 610–623. Springer-Verlag.

Vecchio, D. D., Murray, R. M., and Perona, P. (2003). Decomposition of human motion into dynamics-based primitives with application to drawing tasks. *Automatica*, 39(12):2085–2098.

Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *CVPR*, pages 3169–3176.

Wang, H., Ullah, M. M., Klaser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *Procedings of the British Machine Vision Conference, BMVC*.

Wang, S. B., Quattoni, A., Morency, L.-P., Demirdjian, D., and Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1521–1527.

Wilson, A. D. and Bobick, A. F. (1999). Parametric hidden markov models for gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(9):884–900.

Yang, H., Shao, L., Zheng, F., Wang, L., and Song, Z. (2011). Recent advances and trends in visual tracking: A review. *Neurocomputing*, 74(18):3823–3831.

Yao, B. and Fei-Fei, L. (2010a). Grouplet: A structured image representation for recognizing human and object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Yao, B. and Fei-Fei, L. (2010b). Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Yao, B. and Fei-Fei, L. (2012). Action Recognition with Exemplar Based 2.5D Graph Matching. In *Proceedings of the European Conference on Computer Vision*.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. (2011a). Human action recognition by learning bases of action attributes and parts. In *IEEE International Conference on Computer Vision, ICCV*.

Yao, B., Khosla, A., and Fei-Fei, L. (2011b). Combining randomization and discrimination for fine-grained image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Zacks, J. M. and Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin*, 127(1):3.

Zhou, F., la Torre, F. D., and Hodgins, J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):582–596.